speech

This lecture

- Acoustics.
- Speech production.
- Speech perception.

- Some images from Gray's Anatomy, Jim Glass' course 6.345 (MIT), the Jurafsky & Martin textbook, Encyclopedia Britannica, the Rolling Stones, the Pink Floyds.

UNIVERSITY OF TORONTO
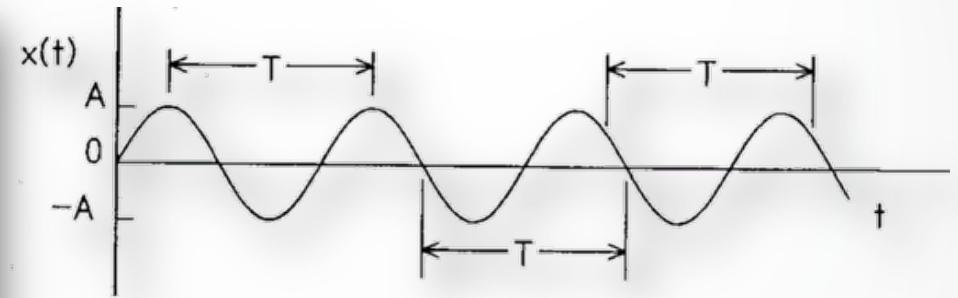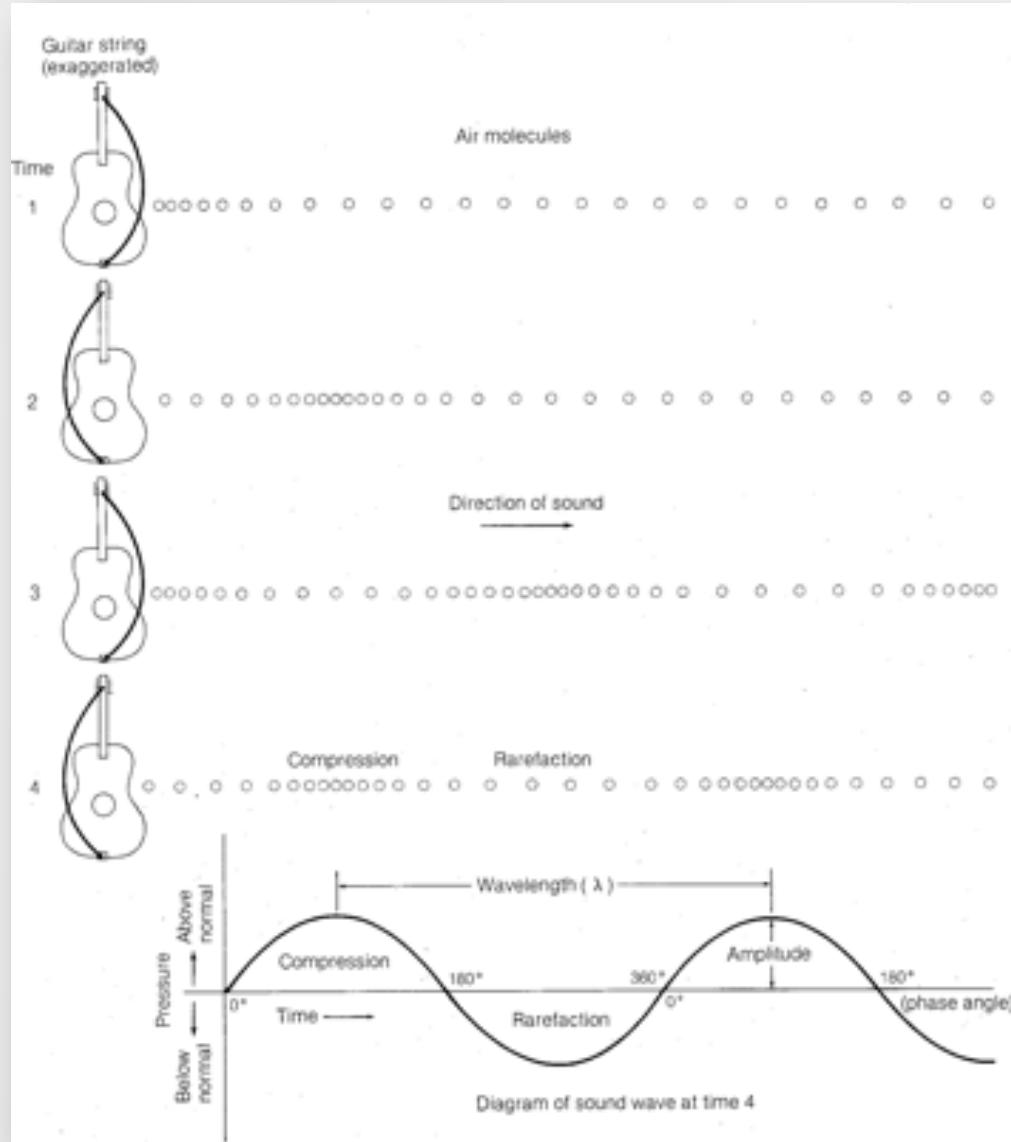
acoustics

What is sound?
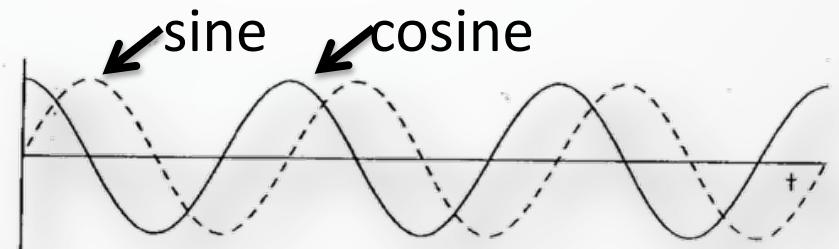
- **Sound** is a time-variant pressure wave created by a vibration.
  - Air particles **hit** each other, setting others in motion.
    - High pressure ≡ **compressions** in the air (C).
    - Low pressure ≡ **rarefactions** within the air (R).

UNIVERSITY OF TORONTO

What is sound?

Guitar string (exaggerated) — Air molecules — Direction of sound — Compression — Rarefaction — Diagram of sound wave at time 4 — Wavelength (λ) — Amplitude — Pressure (Above normal / Below normal)

$x(t)$

**Frequency** $F = 1/T$

sine   cosine
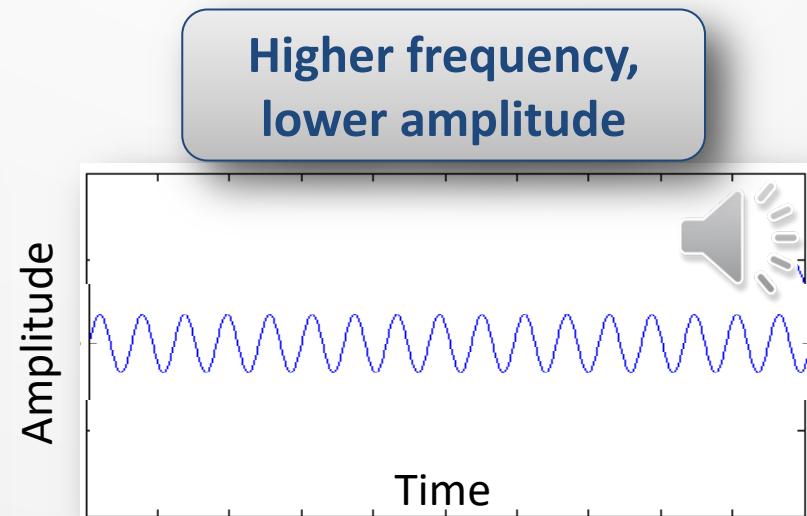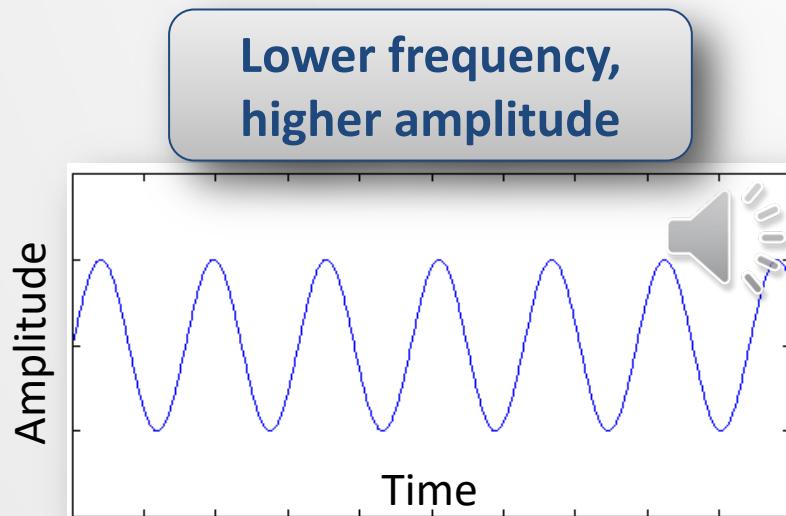
phase ϕ is displacement of a signal in time. E.g., with $\phi = \pi/2$,

$$\sin(x + \phi) = \cos(x)$$

UNIVERSITY OF TORONTO
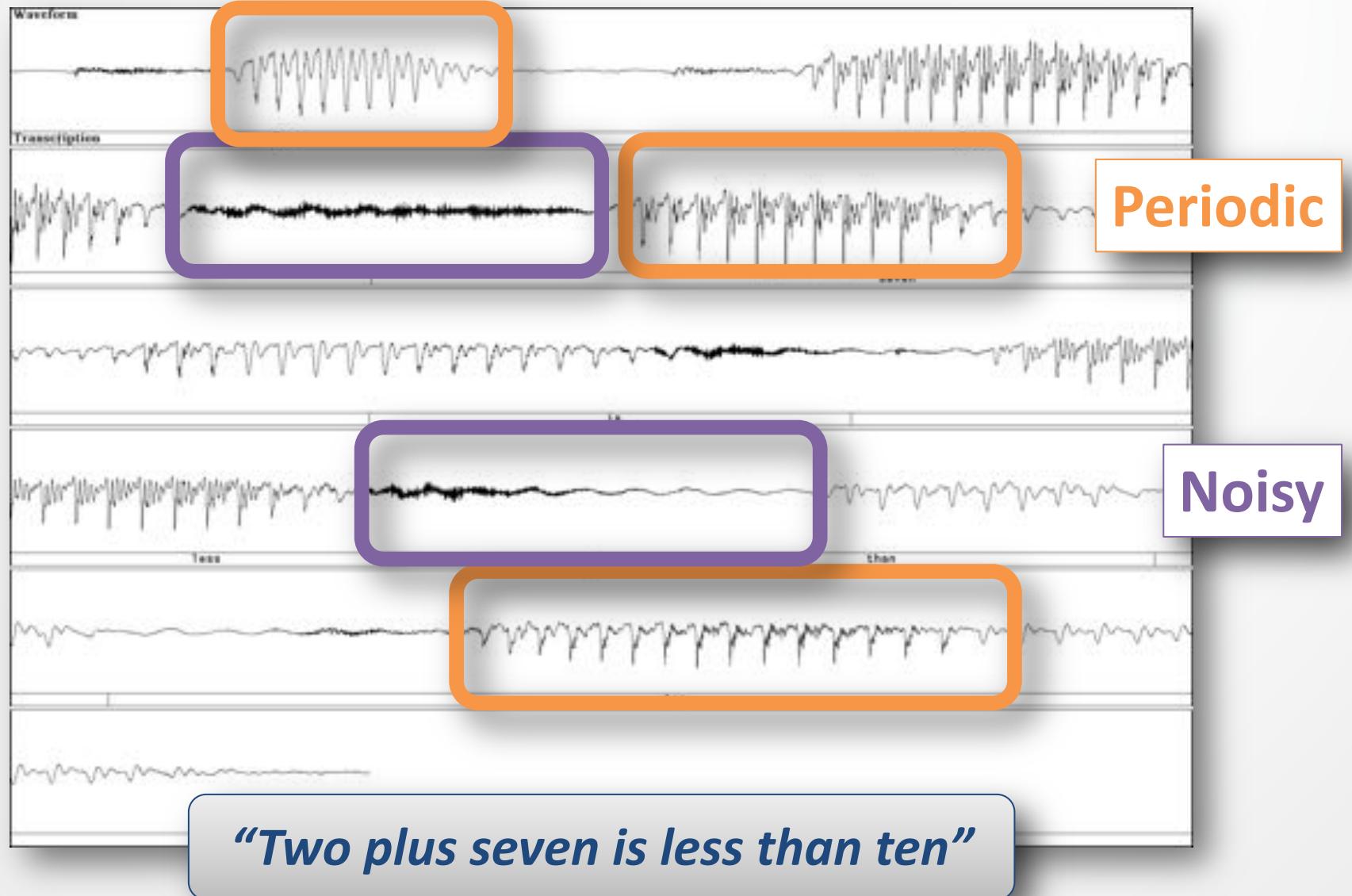
What is sound?

- A single **tone** is a sinusoidal function of pressure and time.
  - **Amplitude**: *n.* The degree of the displacement in the air. This is similar to 'loudness'. Often measured in **Decibels (dB)**.
  - **Frequency**: *n.* The number of cycles within a unit of time. e.g., **1 Hertz (Hz) = 1 oscillation/second**


Lower frequency, higher amplitude

Higher frequency, lower amplitude

UNIVERSITY OF TORONTO

Speech waveforms

Periodic

Noisy

"Two plus seven is less than ten"

UNIVERSITY OF
TORONTO
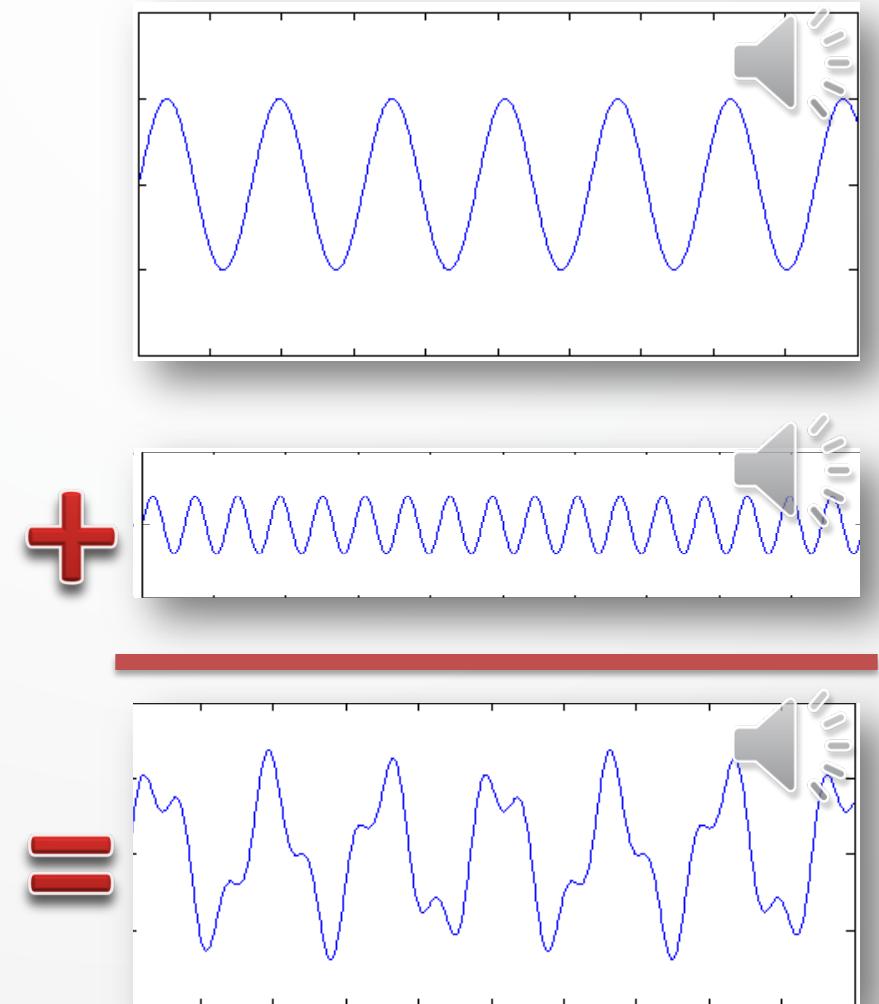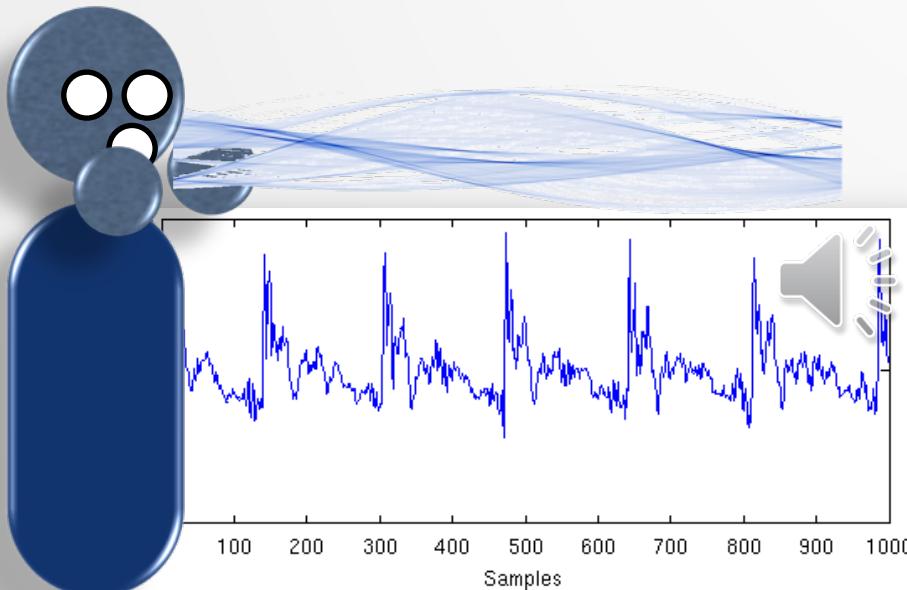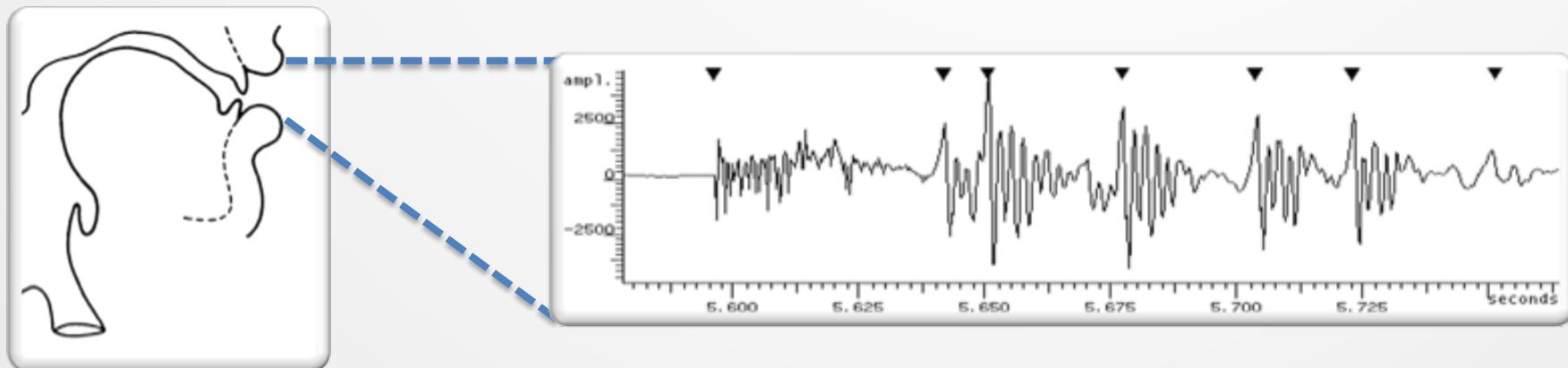
Superposition of sinusoids

- **Superposition**: *n.* the adding of sinusoids together.

- **Phase**: *n.* The horizontal offset of a sinusoid (ϕ).
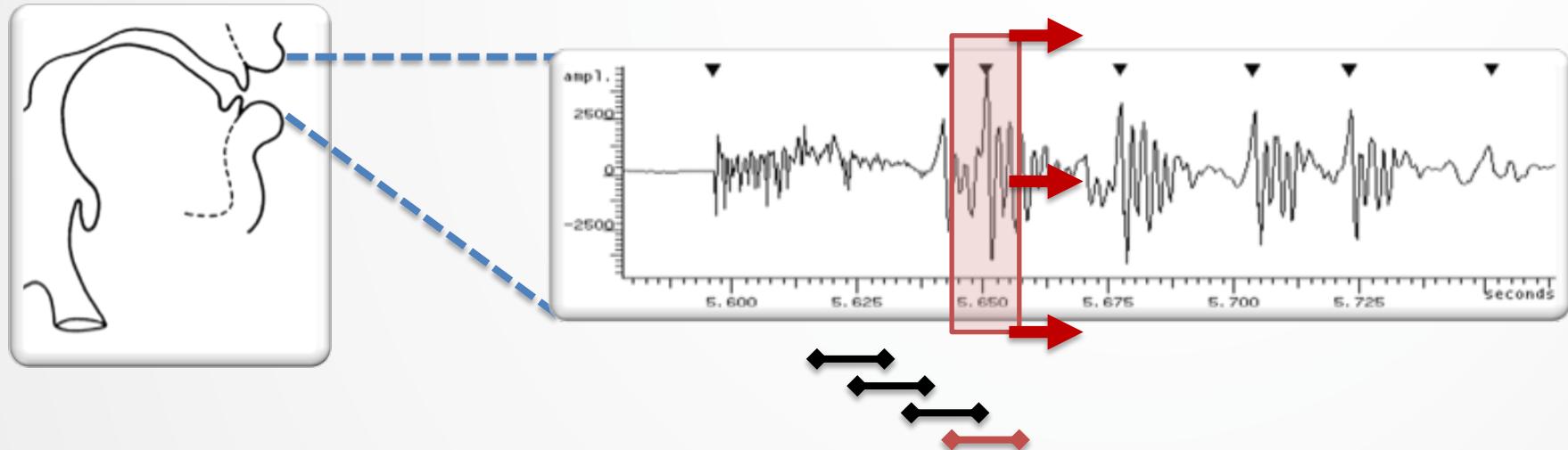
Extracting sinusoids from waveforms

- As we will soon see, the relative **amplitudes** and **frequencies** of the sinusoids that combine in speech are often **extremely indicative** of the **speech units** being uttered.

  - ∴ If we could **separate** the waveform into its component sinusoids, it would help us **classify** the speech being uttered.
  - *But the shape of the signal changes over time*

    (it's not a single repeating pattern)...

UNIVERSITY OF
TORONTO
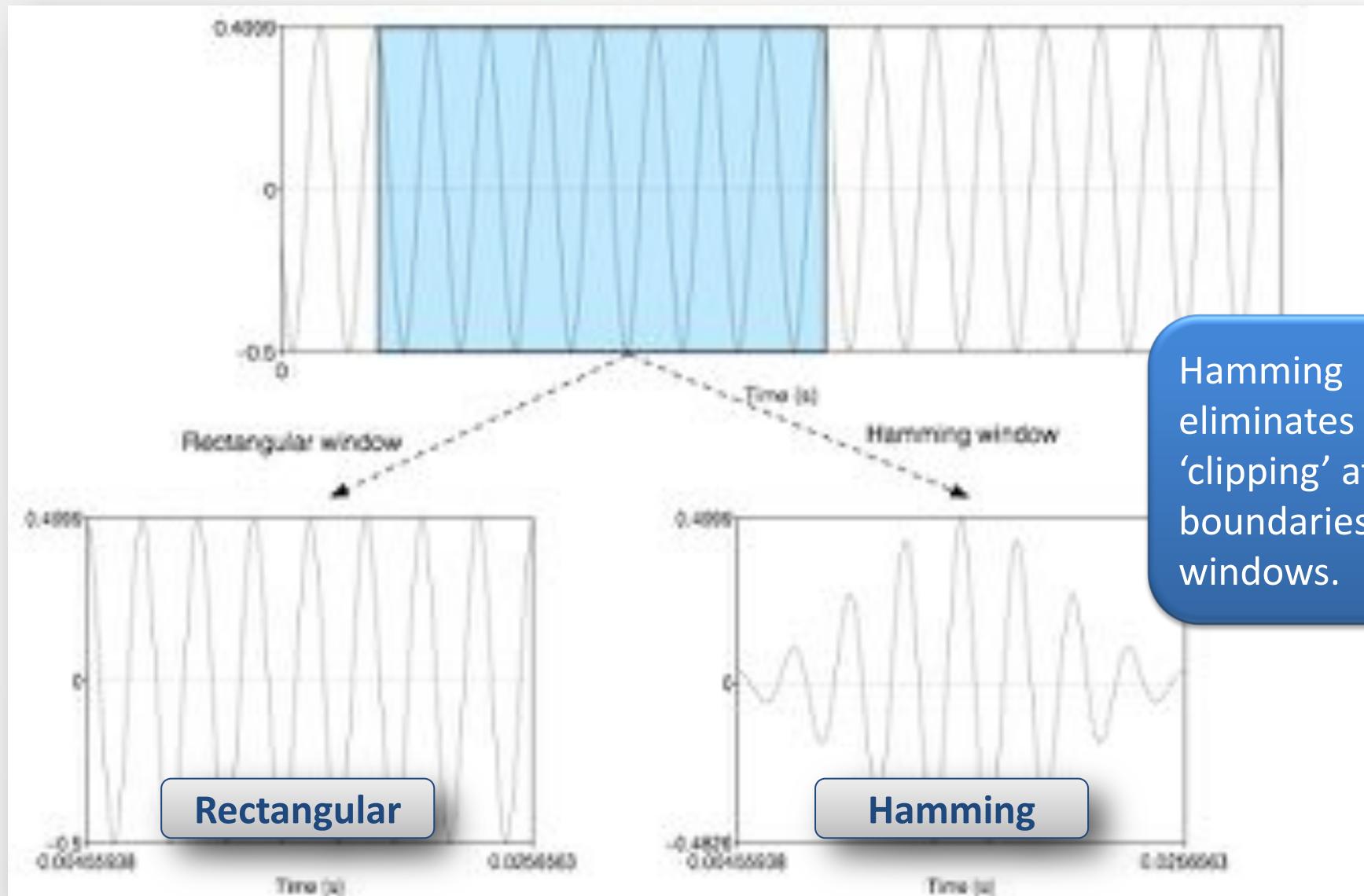
Short-time windowing

Frame

- Speech waveforms change drastically over time.
- We *move* a short analysis window (assumed to be time-invariant) across the waveform in time.
  - E.g. frame shift: 5—10 ms
  - E.g. frame length: 10—25 ms

UNIVERSITY OF
TORONTO

Window types

Rectangular window

Hamming window

Hamming eliminates 'clipping' at the boundaries of windows.

Rectangular

Hamming
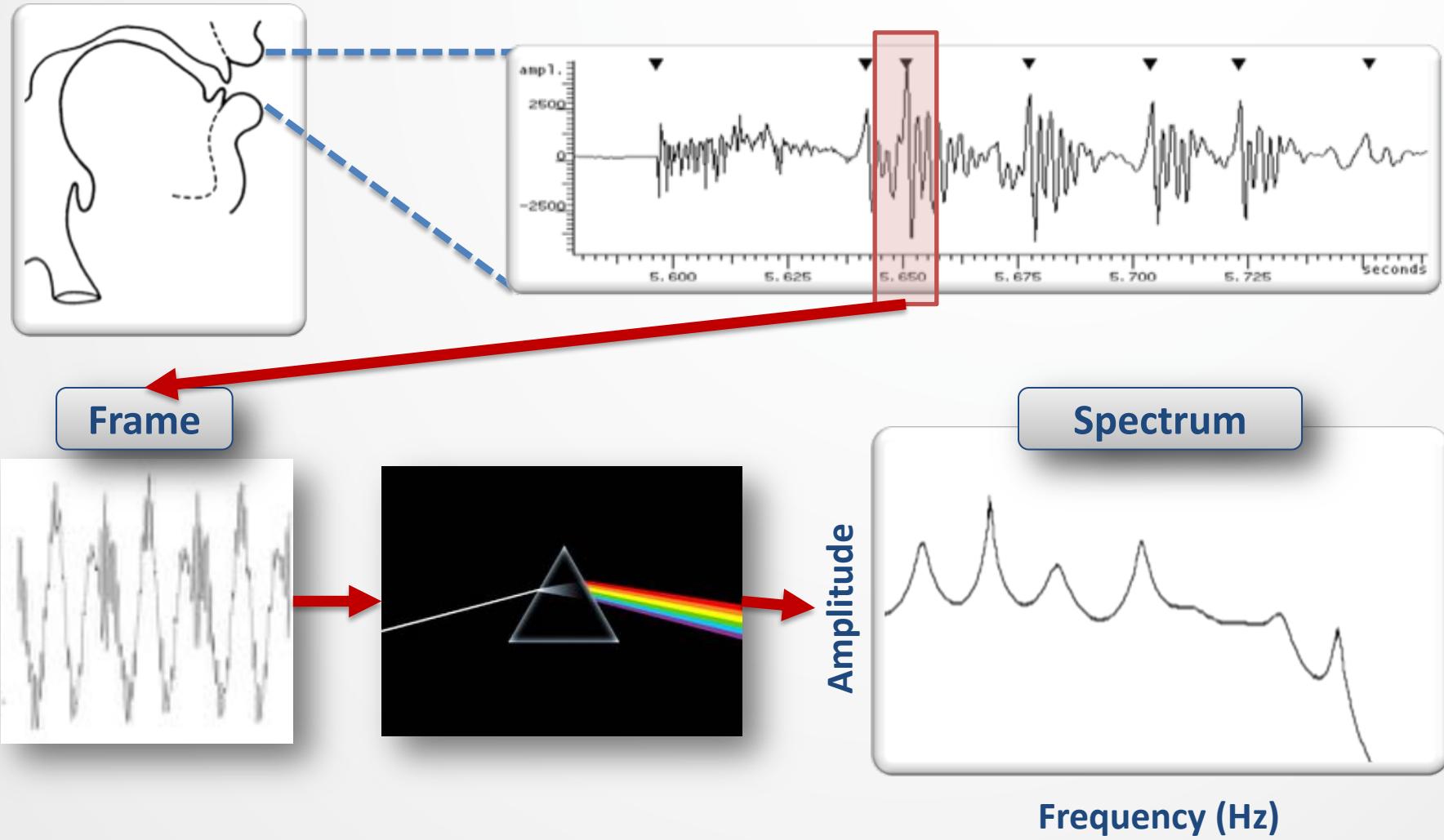
Extracting a spectrum

White light

Any Colour
You Like
(track 8)

UNIVERSITY OF
TORONTO

Extracting a spectrum in a window

Frame

Spectrum

Amplitude

Frequency (Hz)

UNIVERSITY OF
TORONTO
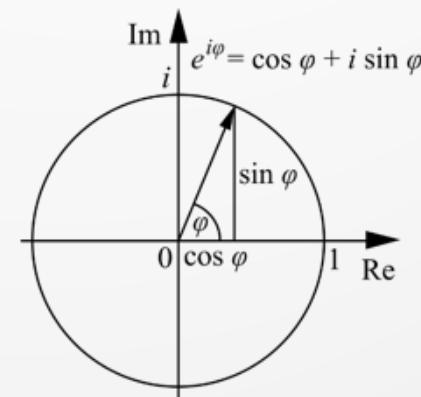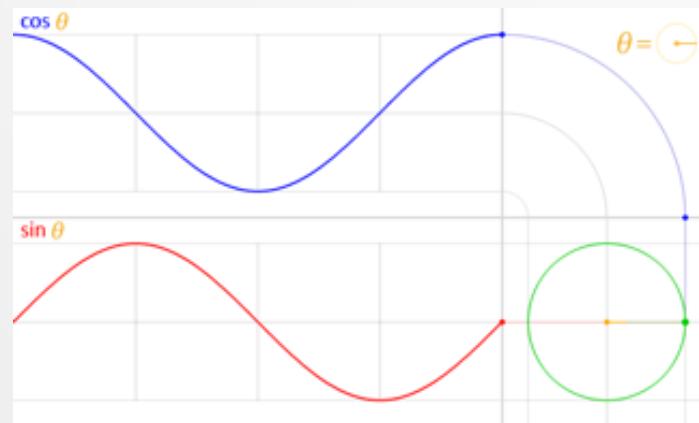
Aside – Euler's formula

- Extracting sinusoids uses a relationship between natural exponent e and sinusoids expressed in **Euler's formula:**

$$e^{ix} = \cos(x) + i\sin(x)$$

$$e^{i\pi} = -1$$

UNIVERSITY OF TORONTO

The continuous Fourier transform

- **Input**: Continuous signal $x(t)$.

- **Output**: Spectrum $X(F)$

$$X(F) = \int_{-\infty}^{\infty} x(t)e^{-i2\pi Ft}\, dt$$

(No need to memorize these)

- It's **invertible**, i.e., $x(t) = \int_{-\infty}^{\infty} X(F)e^{i2\pi Ft}\, dF$.
- It's **linear**, i.e., for $a, b \in \mathbb{C}$,
  if $h(t) = ax(t) + by(t)$,
  then $H(F) = aX(F) + bY(F)$

...
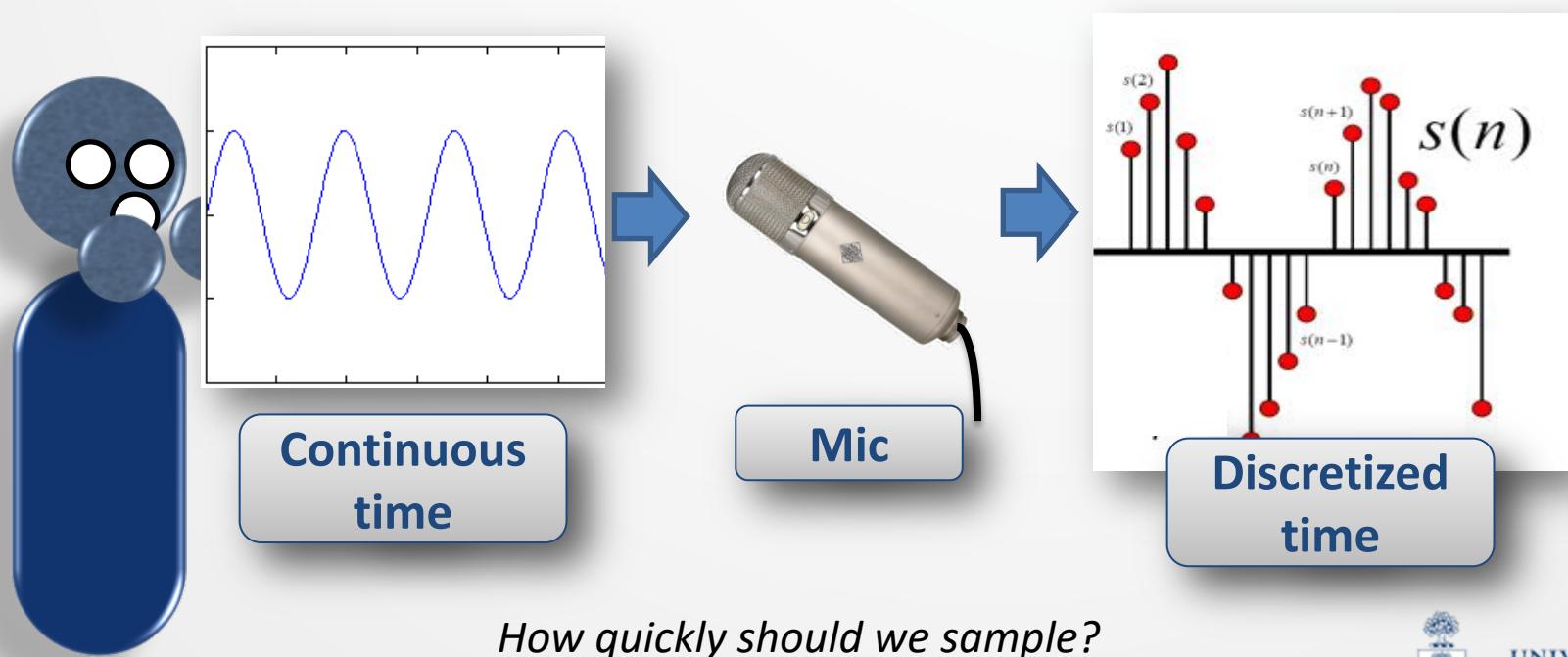
It needs **continuous** input $x(t)$... *uh oh?*

Fun fact: Fourier instructed Champollion.

UNIVERSITY OF TORONTO

# Discrete signal representation
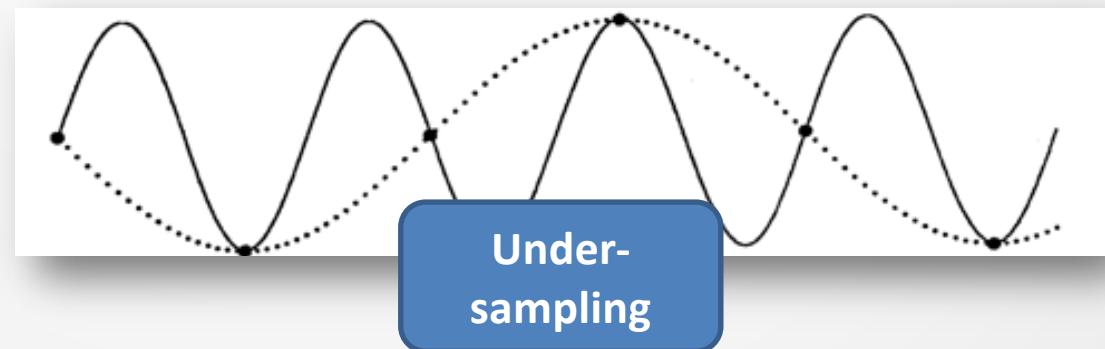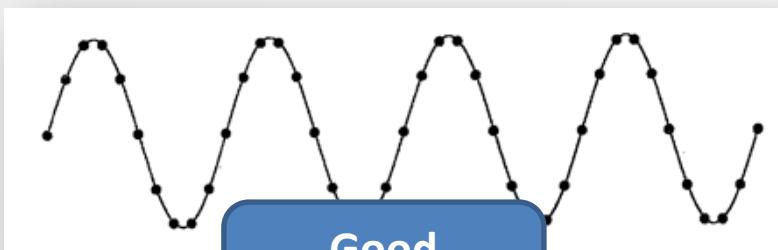
- **Sampling**: *vbg.* measuring the amplitude of a signal at regular intervals.
    - e.g., 44.1 kHz (*CD*), 8 kHz (*telephone*).
    - These amplitudes are initially measured as **continuous** values at **discrete** time steps.

| Continuous time | Mic | Discretized time |

How quickly should we sample?

UNIVERSITY OF TORONTO

Discrete signal representation

- **Nyquist rate**:   *n.* the **minimum** sampling rate necessary to preserve a signal's **maximum** frequency.
  - i.e., **twice** the maximum frequency, since we need ≥ 2 samples/cycle.
  - Human speech is very informative ≤ 4 kHz, $\therefore$ 8 kHz sampling.

Good sampling

Under-sampling

UNIVERSITY OF TORONTO

Discrete Fourier transform (DFT)

- **Input:** Windowed signal $x[0] \dots x[N-1]$.

- **Output:** N complex numbers $X[k]$ $(k \in \mathbb{Z})$
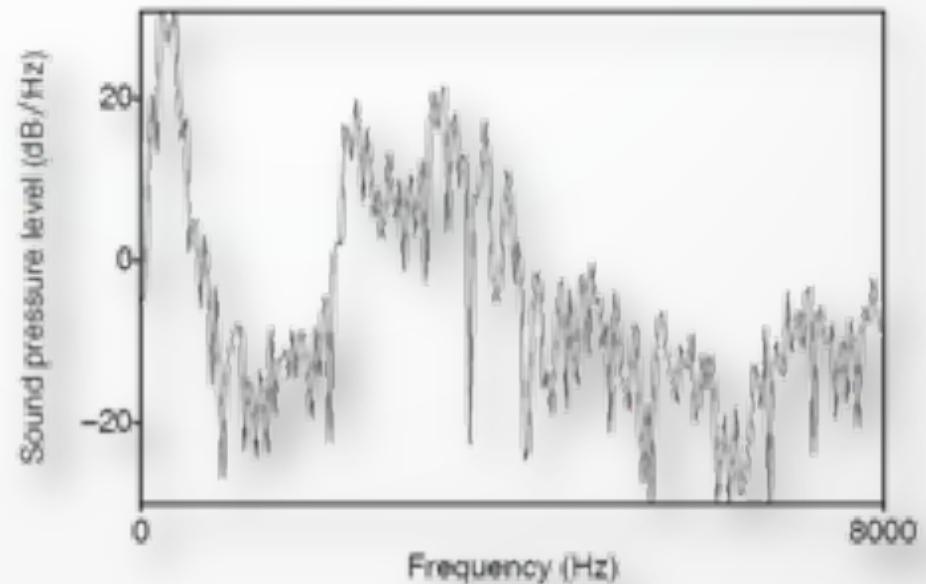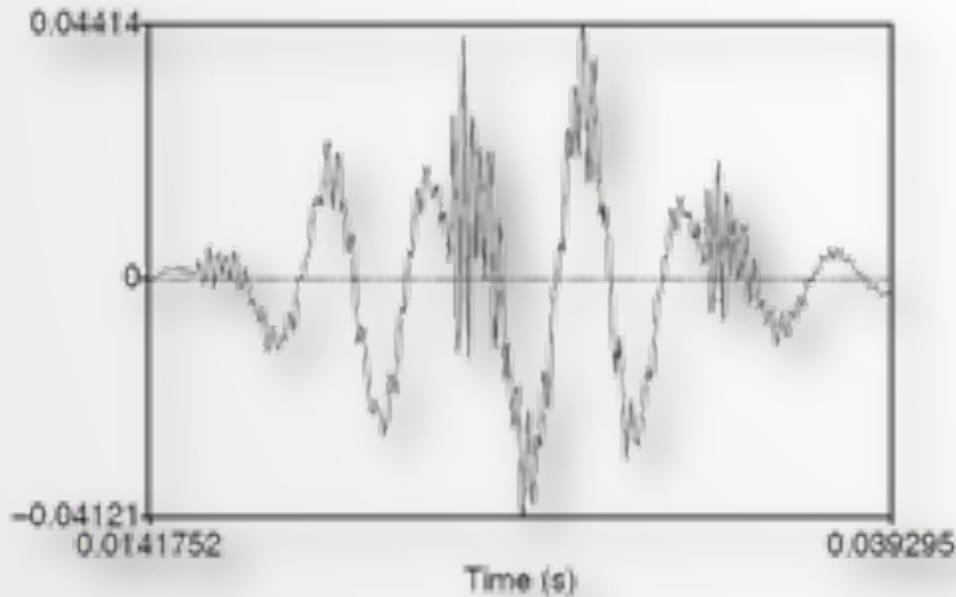
  (No need to memorize these)

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-i2\pi k \frac{n}{N}}$$

- **Algorithm(s):** the **Fast Fourier Transform** (FFT) with complexity $O(N \log N)$.
  - (Aside) The **Cooley-Tukey algorithm** *divides-and-conquers* by breaking the DFT into smaller ones $N = N_1 N_2$.

UNIVERSITY OF TORONTO

Discrete Fourier transform (DFT)

- Below is a 25 ms Hamming-windowed signal from /iy/ as in 'bull sh*ee*p', and its spectrum as computed by the DFT.
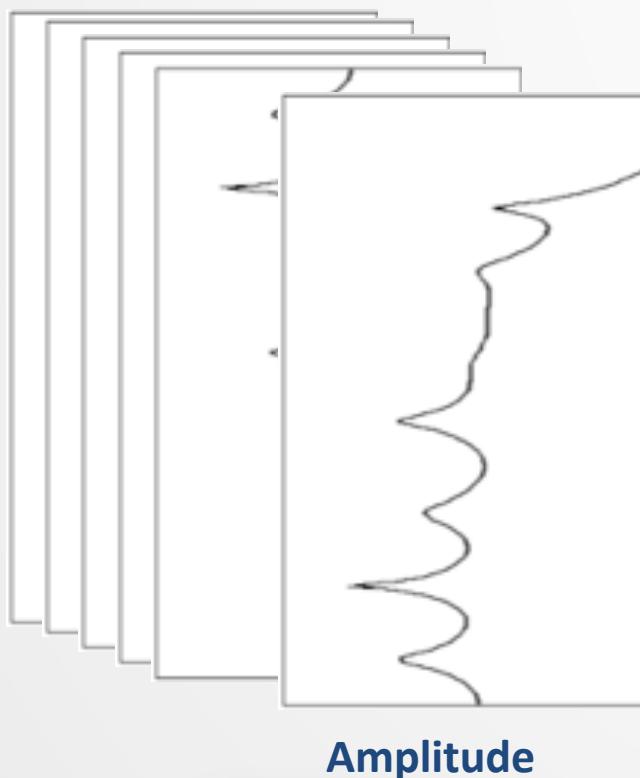
Recall: the Fourier transform is invertible

But this is all just for a small window…

UNIVERSITY OF TORONTO
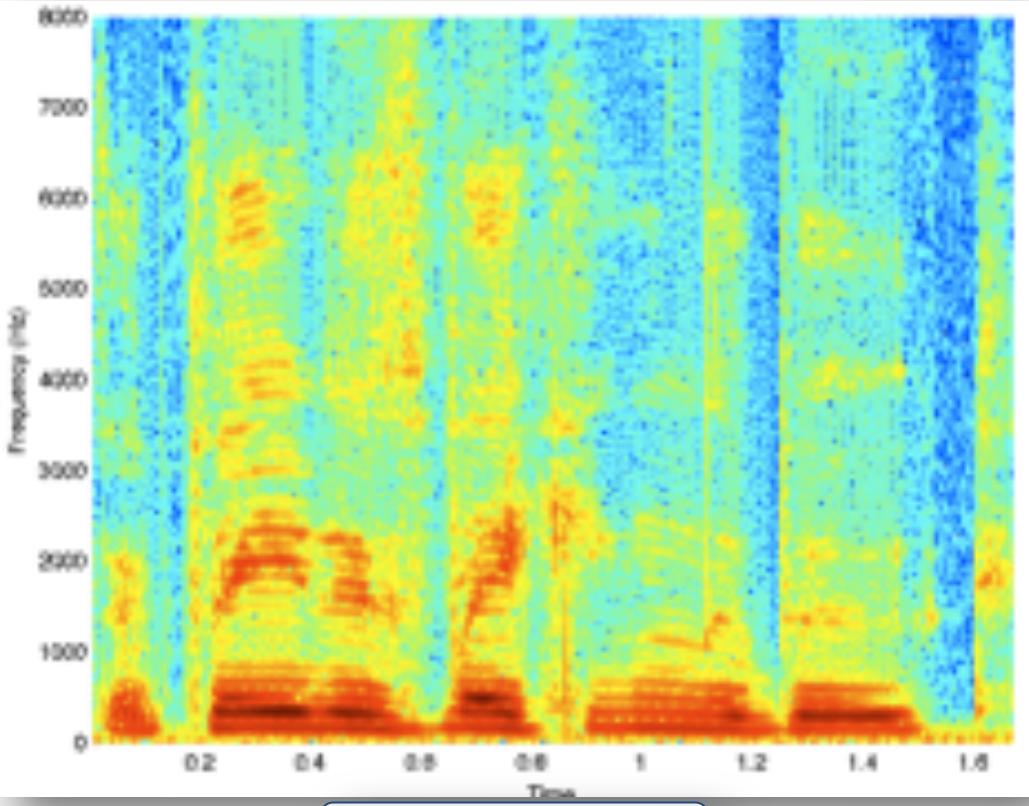
Spectrograms

- **Spectrogram**: *n.* a 3D plot of **amplitude** and **frequency** over **time** (higher 'redness' → higher amplitude).

Frequency (Hz)

Amplitude

Frames

Spectrogram

UNIVERSITY OF
TORONTO
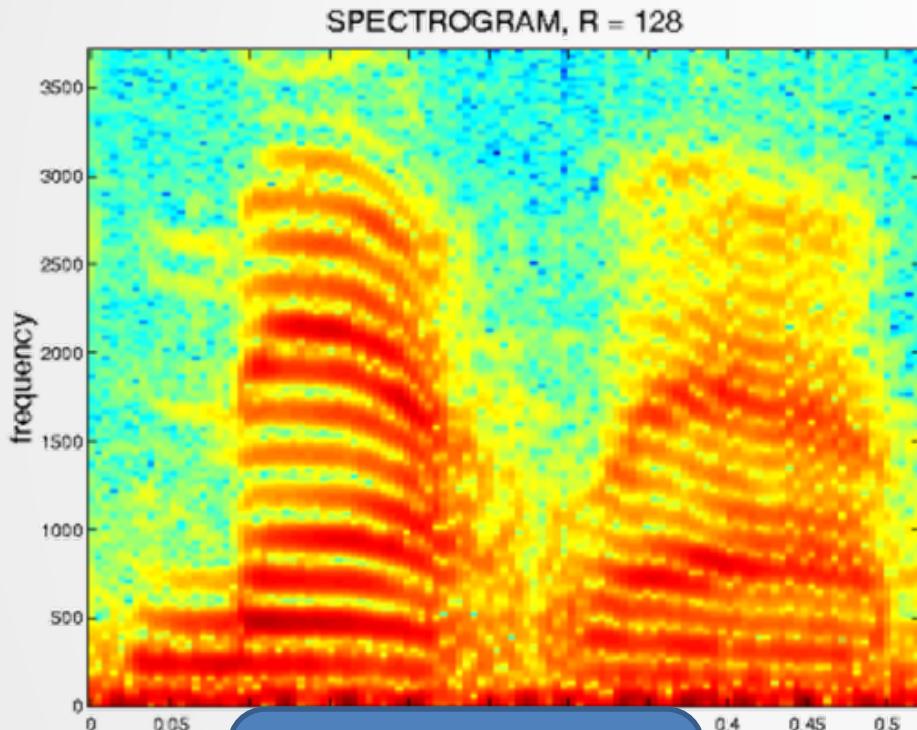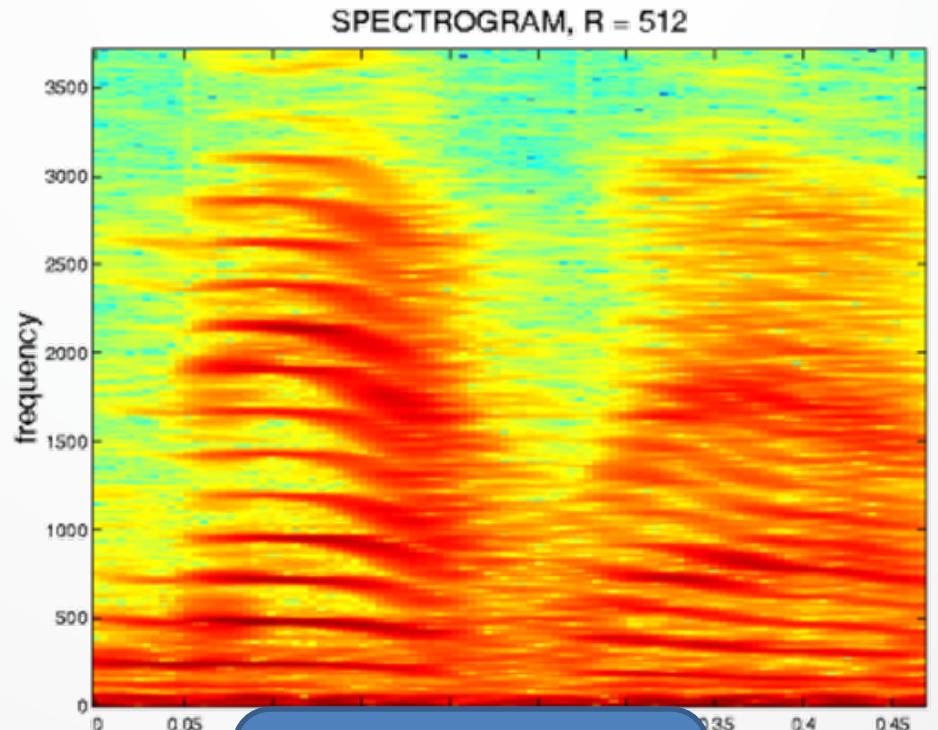
Effect of window length
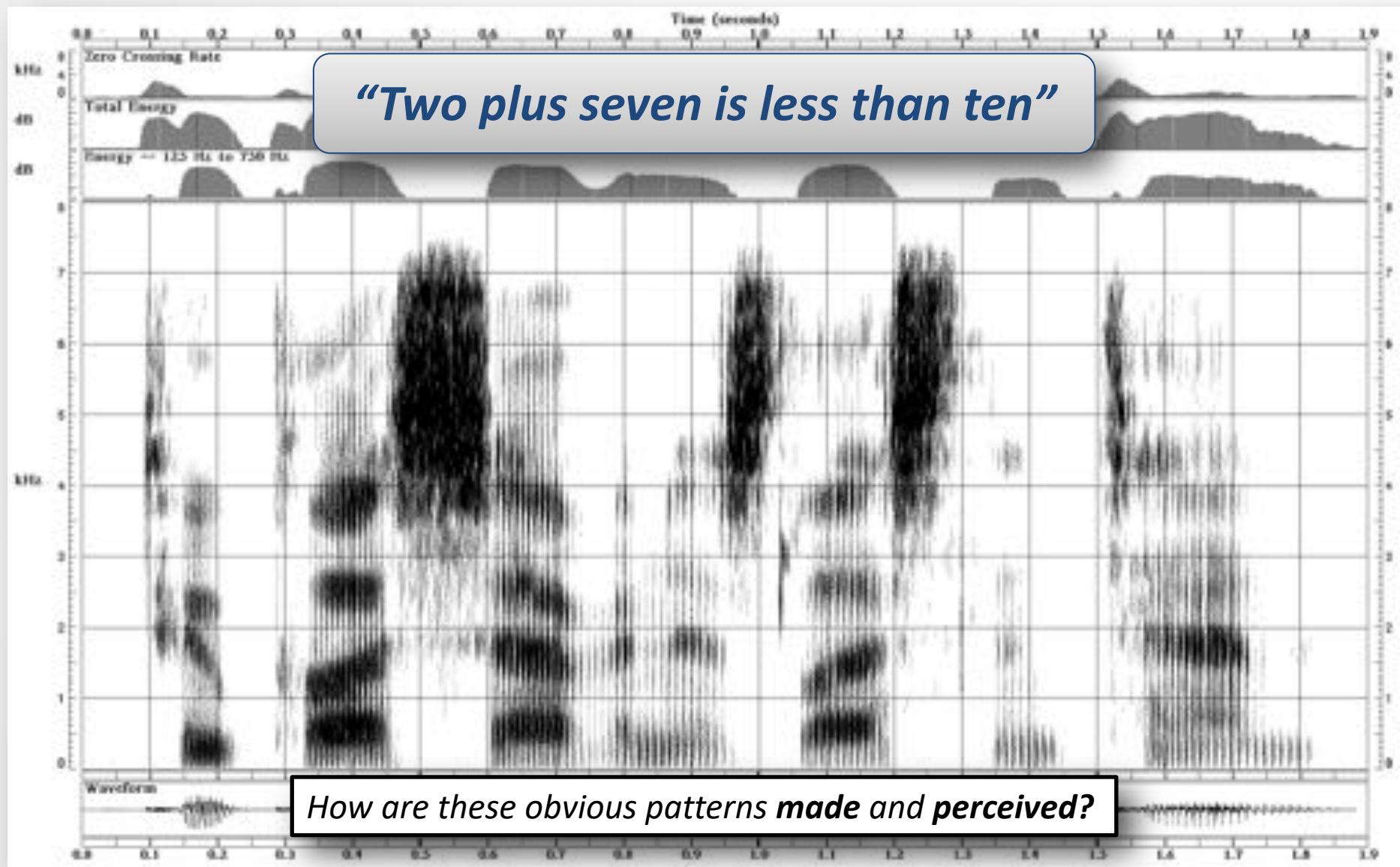
SPECTROGRAM, R = 128

Wide-band (better time resolution)
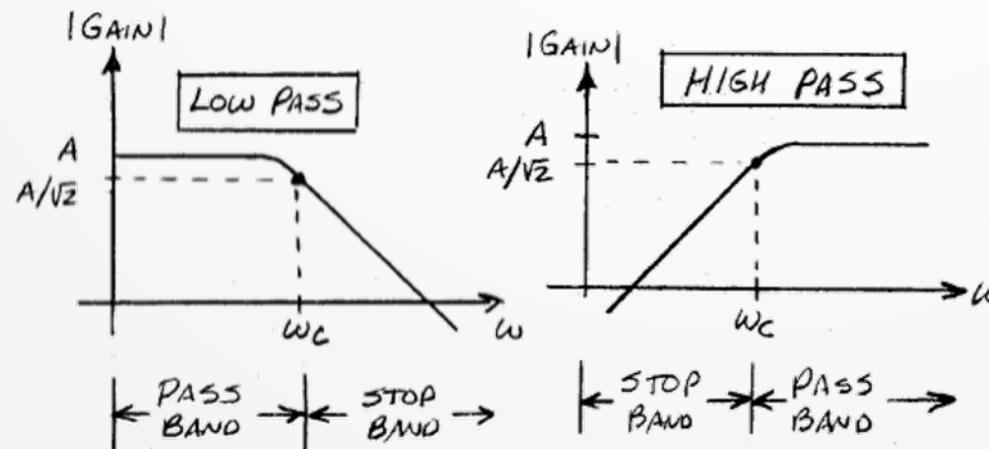
SPECTROGRAM, R = 512

Narrow-band (better frequency resolution)

UNIVERSITY OF
TORONTO

Spectrograms

"Two plus seven is less than ten"

How are these obvious patterns **made** and **perceived?**

Aside – Filtering

- Sometimes you only want **part** of a signal.
 - E.g., you have measurements of lip aperture over time – you know that they can't move > 5-10 Hz.
  - E.g., you know there's some low-frequency Gaussian noise in either the environment or transmission medium.
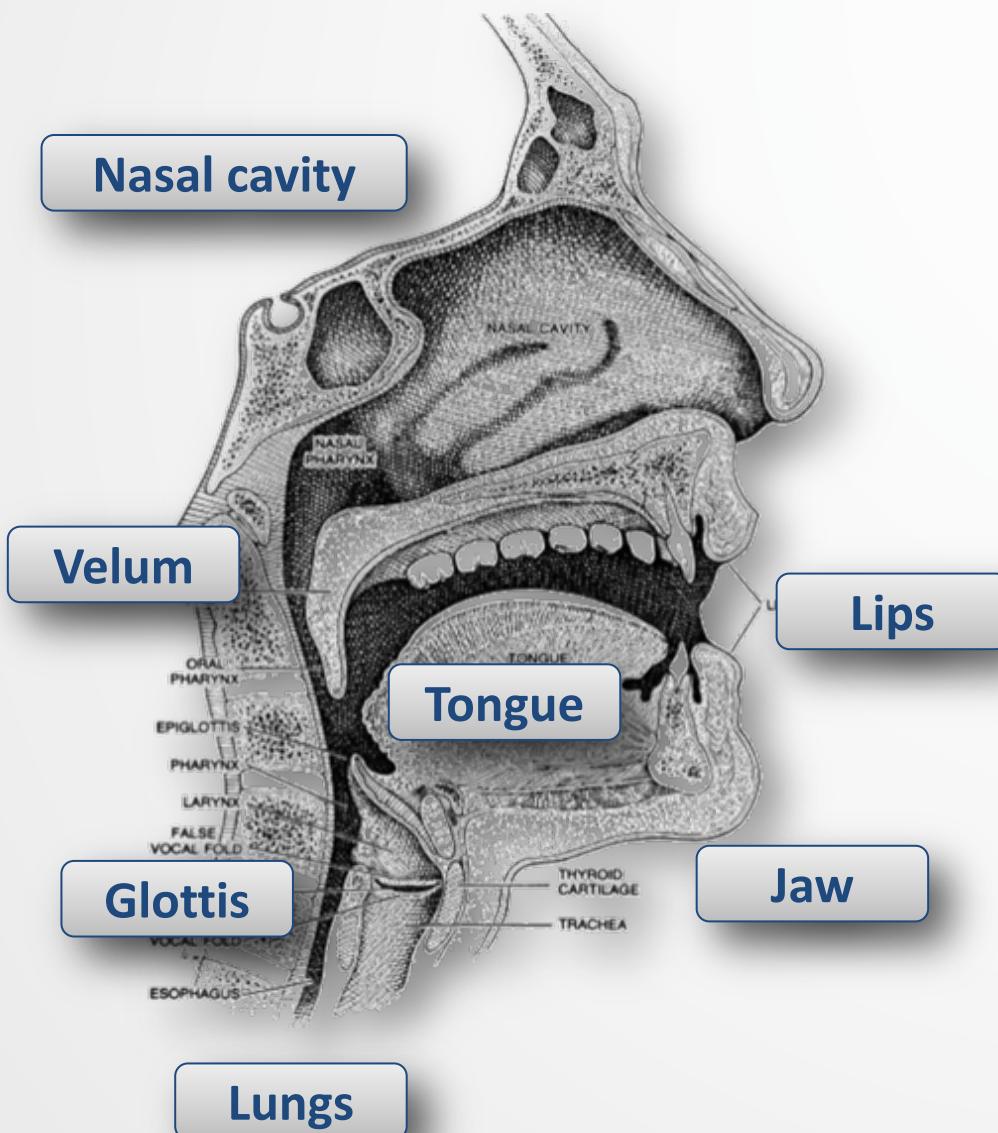
- Low- and high-pass filters can be combined in series, yielding a **band-pass** filter.

speech production

UNIVERSITY OF
TORONTO

The vocal tract

Nasal cavity

Velum

Lips

Tongue

Glottis

Jaw

Lungs
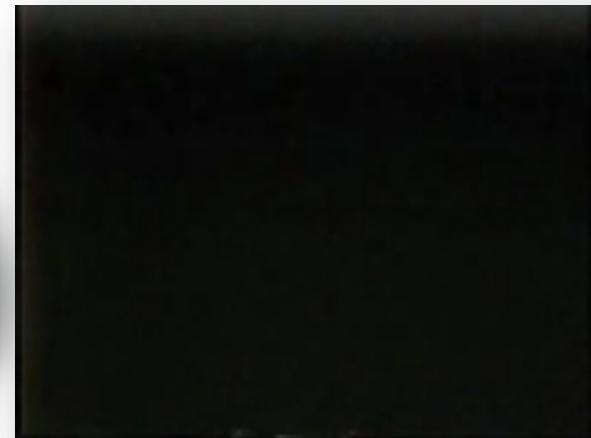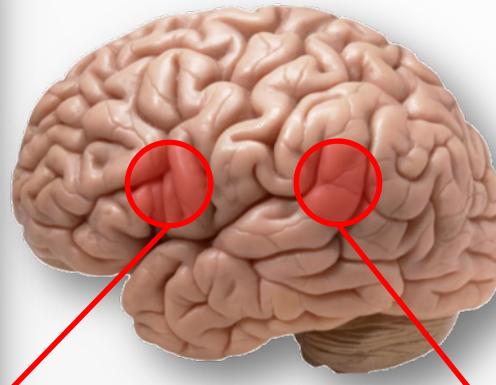
- Many physical structures are co-ordinated in the production of speech.

- Generally, sound is **generated** by passing air through the vocal tract.

- Sound is **modified** by constricting airflow in particular ways.

UNIVERSITY OF
TORONTO

The neurological origins of speech

- Studying how systems break down can indicate how they work.

Broca's aphasia
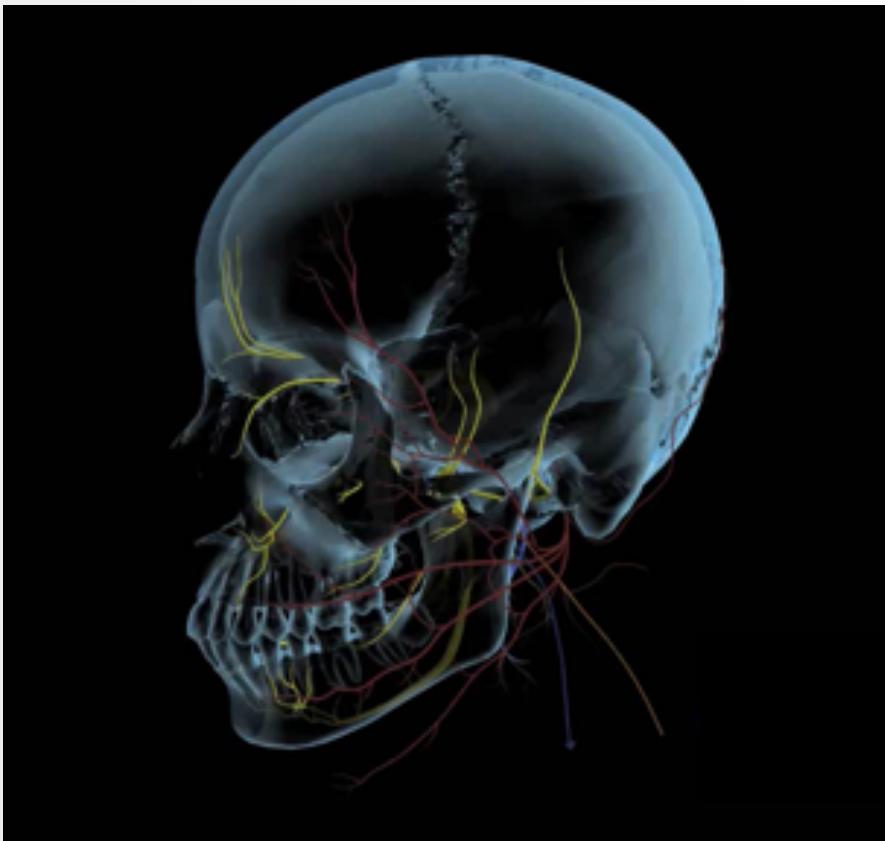
Wernicke's aphasia

- **Reduced** hierarchical **syntax**.
- **Anomia**.
- **Reduced** "mirroring" between **observation** and **execution**.

- **Normal** intonation/rhythm.
- **Meaningless** words.
- '**Jumbled**' syntax.
- **Reduced** comprehension.

UNIVERSITY OF TORONTO

The neurological origins of speech
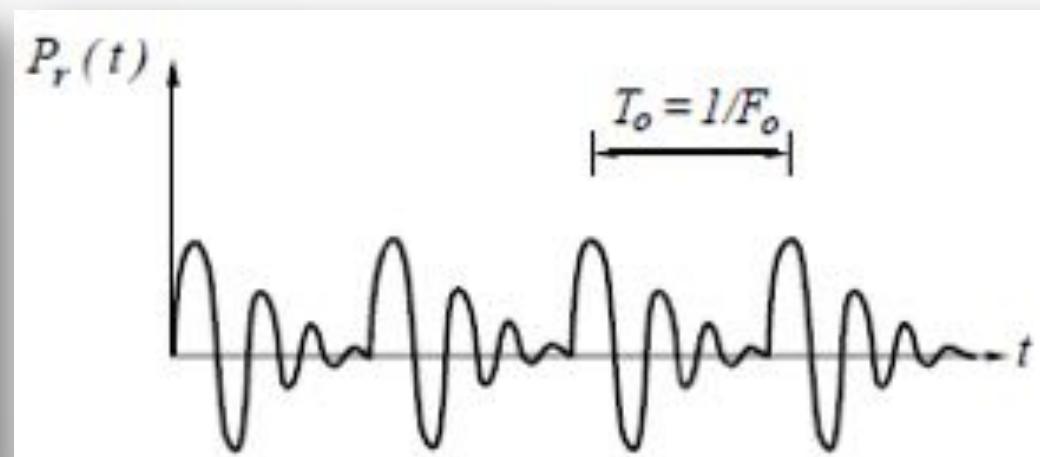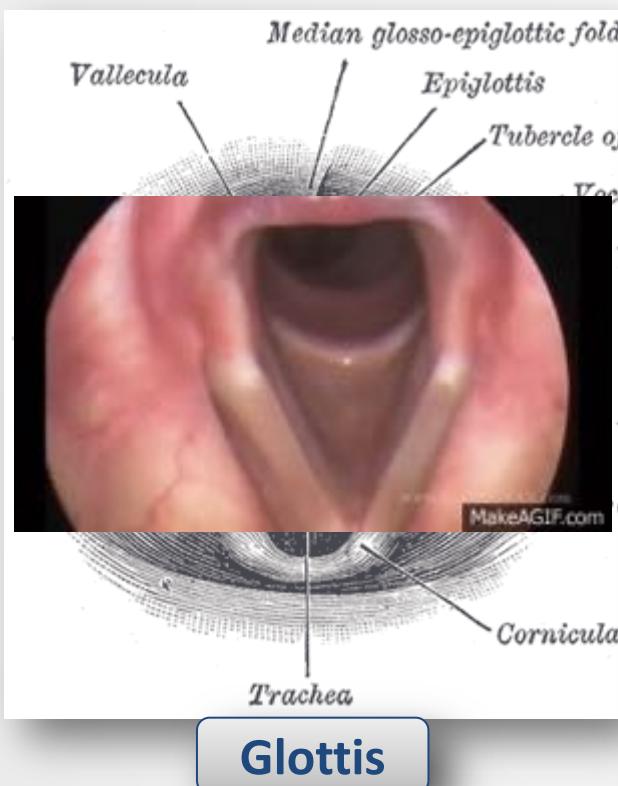
- Cranial nerves carry messages from the brain to the various **articulators.**

- Cranial nerves carry messages from the brain to the various **articulators.**
  - **Damage** to these nerves can result in **neuro-motor** disorders such as cerebral palsy.
  - These may be another example of the noisy channel.

UNIVERSITY OF TORONTO

Fundamental frequency

- F_0: *n.* (**fundamental frequency**), the rate of vibration of the **glottis** – often very **indicative** of the speaker.
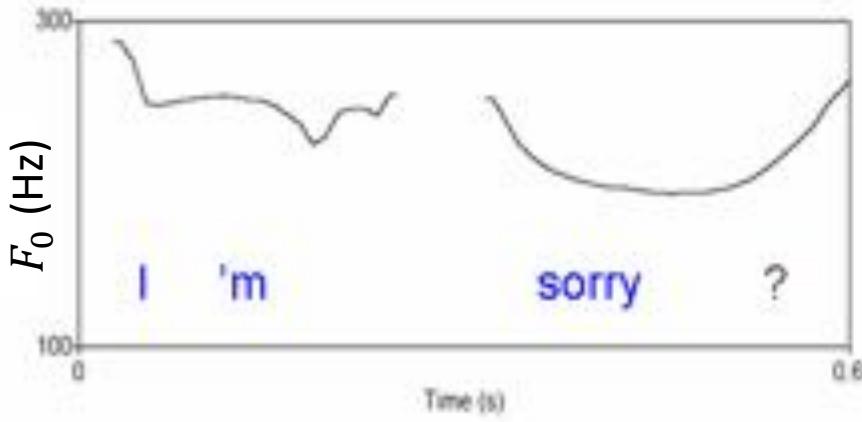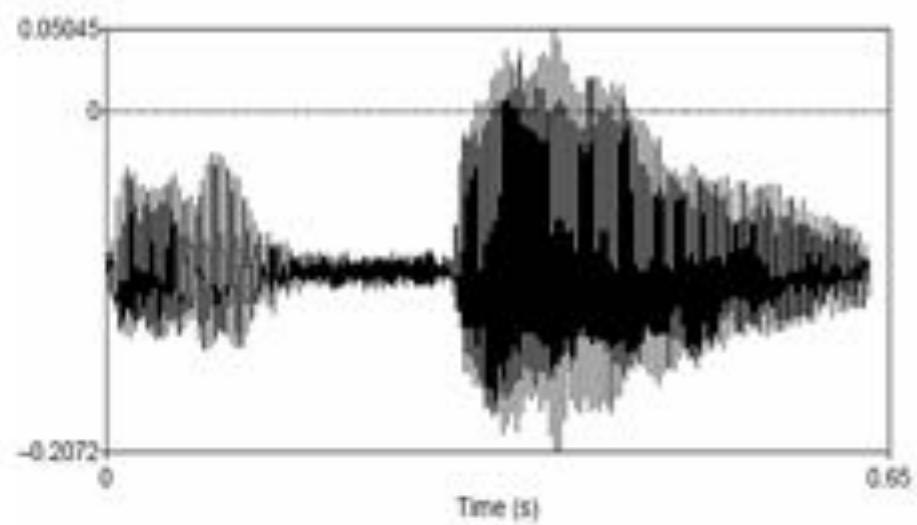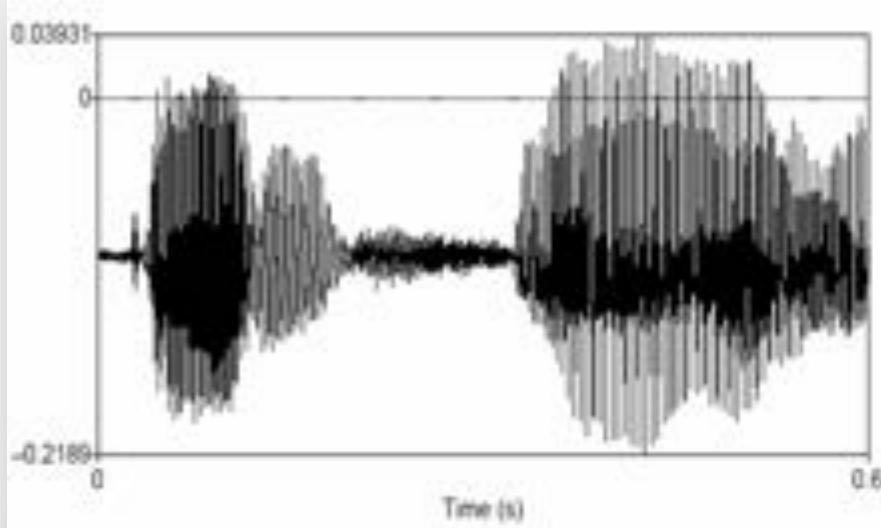
Glottis

	Avg F_0 (Hz)	Min F_0 (Hz)	Max F_0 (Hz)
Men	125	80	200
Women	225	150	350
Children	300	200	500

UNIVERSITY OF TORONTO
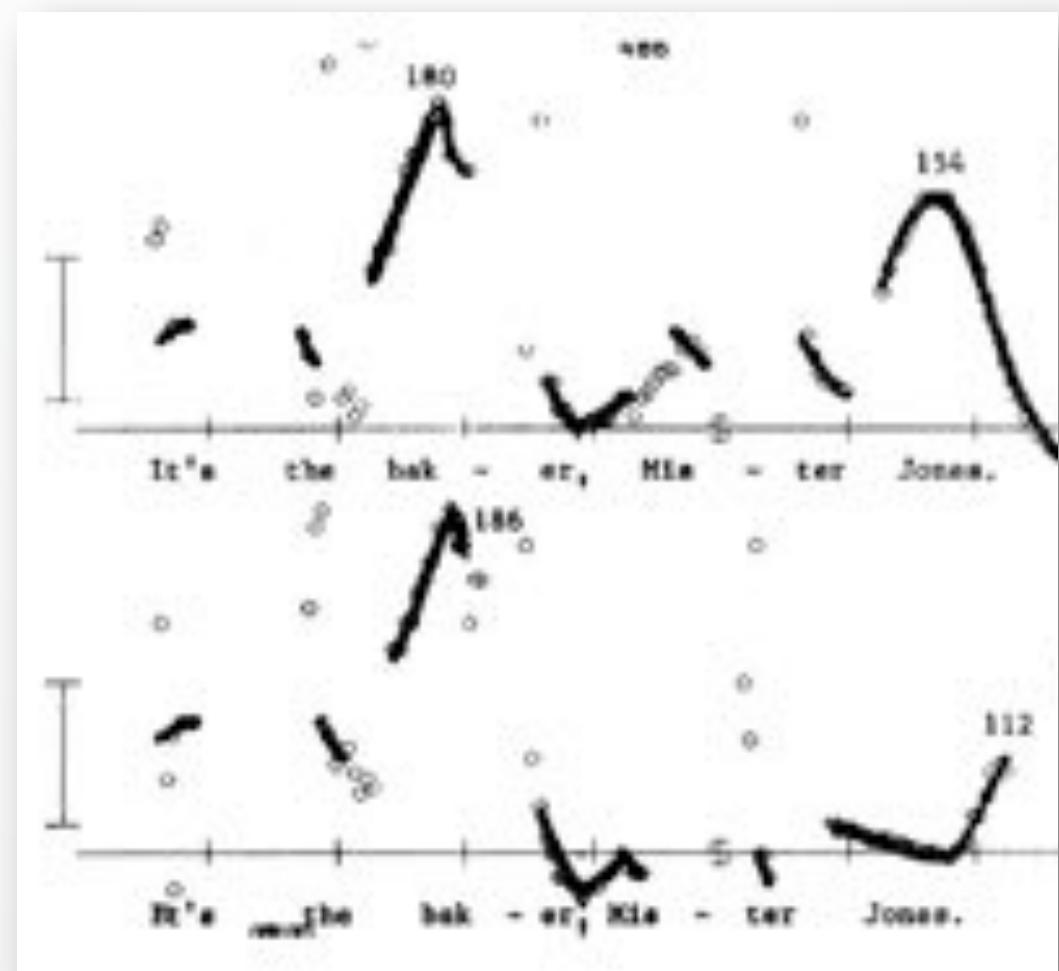
Prosody

- **Sonorant**: *n.* Any **sustained** sound in which the **glottis** is vibrating (i.e., the sound is '**voiced**').
  - Includes some consonants (e.g., /w/, /m/).

- **Prosody**: *n.* the **modification** of speech acoustics in order to convey some **extra-lexical** meaning:
  - **Pitch**: Changing of F_0 over time.
  - **Duration**: The length in time of sonorants.
  - **Loudness**: The amount of **energy** produced by the **lungs**.

UNIVERSITY OF TORONTO

Pitch prosody example

Pitch can modify meaning

- e.g., I ask you
  "*who* is that?"

- e.g., I ask you
  "*what is his job?*"

Pitch tends to rise when uttering novel or important information.

UNIVERSITY OF TORONTO

Pitch can modify meaning
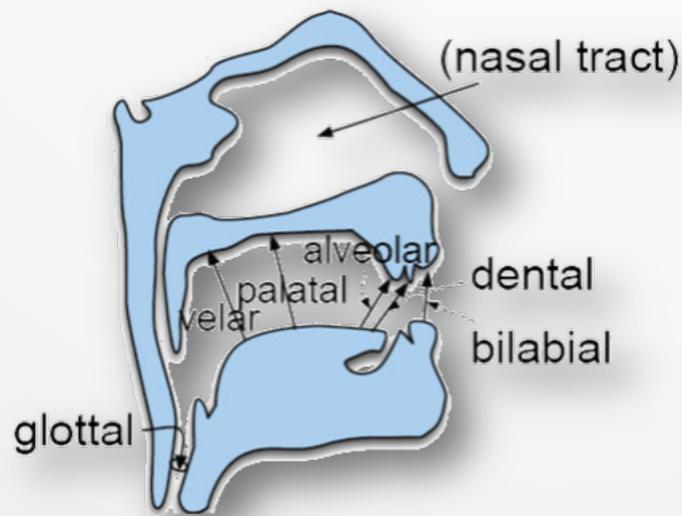
- ***I*** never said she stole my money. (Someone else said it)
- I ***never*** said she stole my money. (It never happened)
- I never ***said*** she stole my money. (I just hinted at it)
- I never said ***she*** stole my money. (Someone else stole it)
- I never said she ***stole*** my money. (She just borrowed it)
- I never said she stole ***my*** money. (She stole someone else's)
- I never said she stole my ***money***. (She stole my heart).

UNIVERSITY OF TORONTO

Phonemes

- **Phoneme**: *n.* a distinctive unit of speech sound.
- Phonemes can be partitioned into **manners of articulation**:
    - **Vowels**: **open** vocal tract, no nasal air.
    - **Fricatives**: **noisy**, with air passing through a tight constriction (e.g., '*shift*').
    - **Stops/plosives**: **complete** vocal tract constriction and burst of energy (e.g., '*papa*').
    - **Nasals**: air passes through the **nasal** cavity (e.g., '*mama*').
    - **Semivowels**: similar to vowels, but typically with more constriction (e.g., '*wall*').
    - **Affricates**: Alveolar stop followed by fricative.

Place of articulation

- The **location** of the *primary constriction* can be:
  - **Alveolar**:   constriction near the alveolar ridge (e.g., /*t*/)
  - **Bilabial**:   touching of the lips together (e.g., /*m*/, /*p*/)
  - **Dental**:   constriction of/at the teeth (e.g., /*th*/)
  - **Labiodental**: constriction between lip and teeth (e.g., /*f*/)
  - **Velar**:   constriction at or near the velum (e.g., /*k*/).

UNIVERSITY OF TORONTO

Phonemic alphabets

- There are several alphabets that categorize the sounds of speech.
  - The **International Phonetic Alphabet (IPA)** is popular, but it uses non-ASCII symbols.
  - The **TIMIT** phonemic alphabet will be used by **default** in this course.

  - Other popular alphabets include **ARPAbet**, **Worldbet**, and **OGIbet**, usually adding special cases.
    - E.g., /pcl/ is the period of silence immediately before a /p/.

TIMIT	IPA	e.g.
/iy/	/iʲ/	b*ea*t
/ih/	/ɪ/	b*i*t
/eh/	/ɛ/	b*e*t
/ae/	/æ/	b*a*t
/aa/	/ɑ/	B*o*b
/ah/	/ʌ/	b*u*t
/ao/	/ɔ/	b*ou*ght
/uh/	/ʊ/	b*oo*k
/uw/	/u/	b*oo*t
/ux/	/ʉ/	s*ui*t
/ax/	/ə/	*a*bout

UNIVERSITY OF TORONTO

TIMIT Phonemic alphabet (incomplete)

Vowel	e.g.
/iy/	b*ea*t
/ih/	b*i*t
/eh/	b*e*t
/ae/	b*a*t
/aa/	B*o*b
/ah/	b*u*t
/ao/	b*ou*ght
/uh/	b*oo*k
/uw/	b*oo*t
/ux/	s*ui*t
/ax/	*a*bout

stop	e.g.
/b/	*Bilb*o
/d/	*d*a*d*a
/g/	*G*a*g*a
/p/	*Pipp*in
/t/	*T*oo*t*s
/k/	*k*i*ck*

nasal	e.g.
/m/	*Ma*m*a*
/n/	*n*oo*n*
/ng/	thi*ng*

fricative	e.g.
/s/	*S*ea
/f/	*F*rank
/z/	*Z*appa
/th/	*th*is
/sh/	*Sh*ip
/zh/	a*z*ure
/v/	*V*ogon
/dh/	*th*en

• • •

(Incomplete)

UNIVERSITY OF TORONTO

Phoneme sequences

- Often, we assume that a **spoken utterance** can be **partitioned** into a **sequence** of **non-overlapping** phonemes.
  - Demarking the periods during which certain phonemes are being uttered is called **transcription** or **annotation** (*).
  - This approach has problems (e.g., when *exactly* does one phoneme end and another begin?), but it's useful for **classification**.

Waveform

This looks periodic

/t/   /ux/

*What are some characteristics of the six **manners** of articulation?*

UNIVERSITY OF TORONTO

Vowels (1/6)

- There are approximately **19** vowels in Canadian English, including **diphthongs** in which the articulators **move** over time.

- Vowels are distinguished primarily by their **formants**. (?)

Mono-phthong	e.g.
/iy/	b<u>ea</u>t
/ih/	b<u>i</u>t
/eh/	b<u>e</u>t
/ae/	b<u>a</u>t
/aa/	B<u>o</u>b
/ao/	b<u>ou</u>ght
/ah/	b<u>u</u>t
/uh/	b<u>oo</u>k
/uw/	b<u>oo</u>t
/ax/	<u>a</u>bout
/ix/	ros<u>e</u>s

diphthong	e.g.
/ey/	b<u>ai</u>t
/ow/	b<u>oa</u>t
/ay/	b<u>i</u>te
/oy/	b<u>oy</u>
/aw/	b<u>ou</u>t
/ux/	s<u>ui</u>t

other	e.g.
/er/	B<u>er</u>t
/axr/	b<u>u</u>tter

UNIVERSITY OF TORONTO

The uniform tube

Closed, vibrating end		Open, radiating end
glottis	17 cm	lips

- The positions of the tongue, jaw, and lips change the **shape** and **cross-sectional** area of the vocal tract.

Vocal-tract area function

Area (cm²)

Distance from larynx (cm)

UNIVERSITY OF
TORONTO
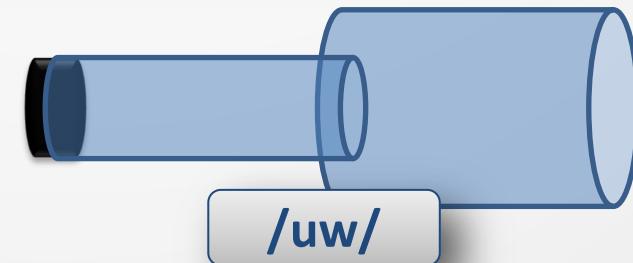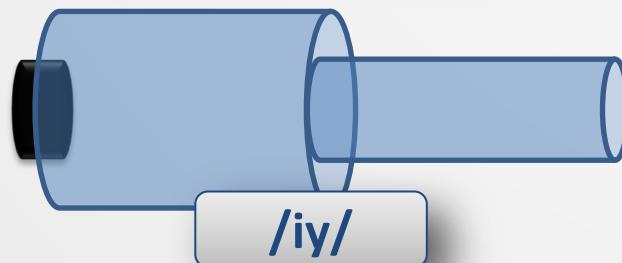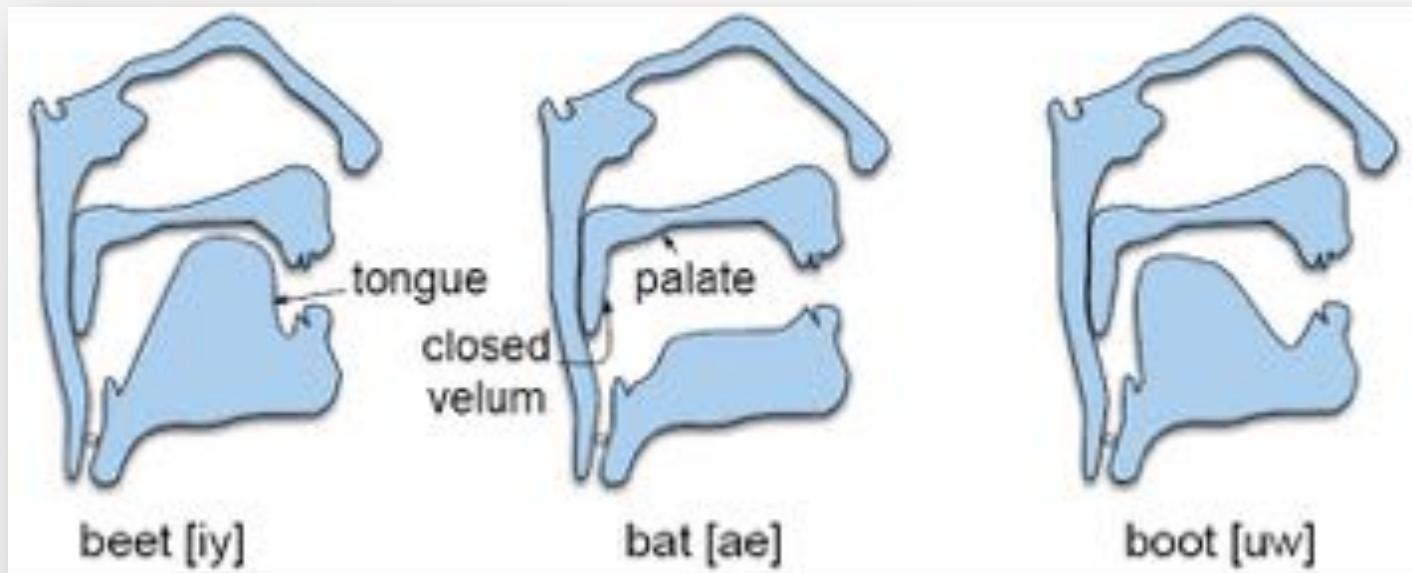
Uniform tubes in practice

- Many **musical instruments** are based on the idea of uniform (or, in many cases, bent) tubes.

- **Longer** tubes produce '**deeper**' sounds (lower frequencies).
  - A tube ½ the length of another will be 1 octave higher.

UNIVERSITY OF
TORONTO

Vowels as concatenated tubes

- The vocal tract can be modelled as the concatenation of dozens, hundreds, or thousands of tubes.

beet [iy]   bat [ae]   boot [uw]

/iy/

/uw/

UNIVERSITY OF
TORONTO

Aside – waves in concatenated tubes

- We model the **volume velocity** U_k and the **pressure variation** p_k at position x in the kth lossless tube (whose area is A_k) at time t

$$U_k(x,t) = U_k^+\left(t - \frac{x}{c}\right) - U_k^-\left(t + \frac{x}{c}\right)$$

$$p_k(x,t) = \frac{\rho c}{A_k}\left[U_k^+\left(t - \frac{x}{c}\right) + U_k^-\left(t + \frac{x}{c}\right)\right]$$

where
c is the speed of sound,
ρ is the density of air.

UNIVERSITY OF TORONTO
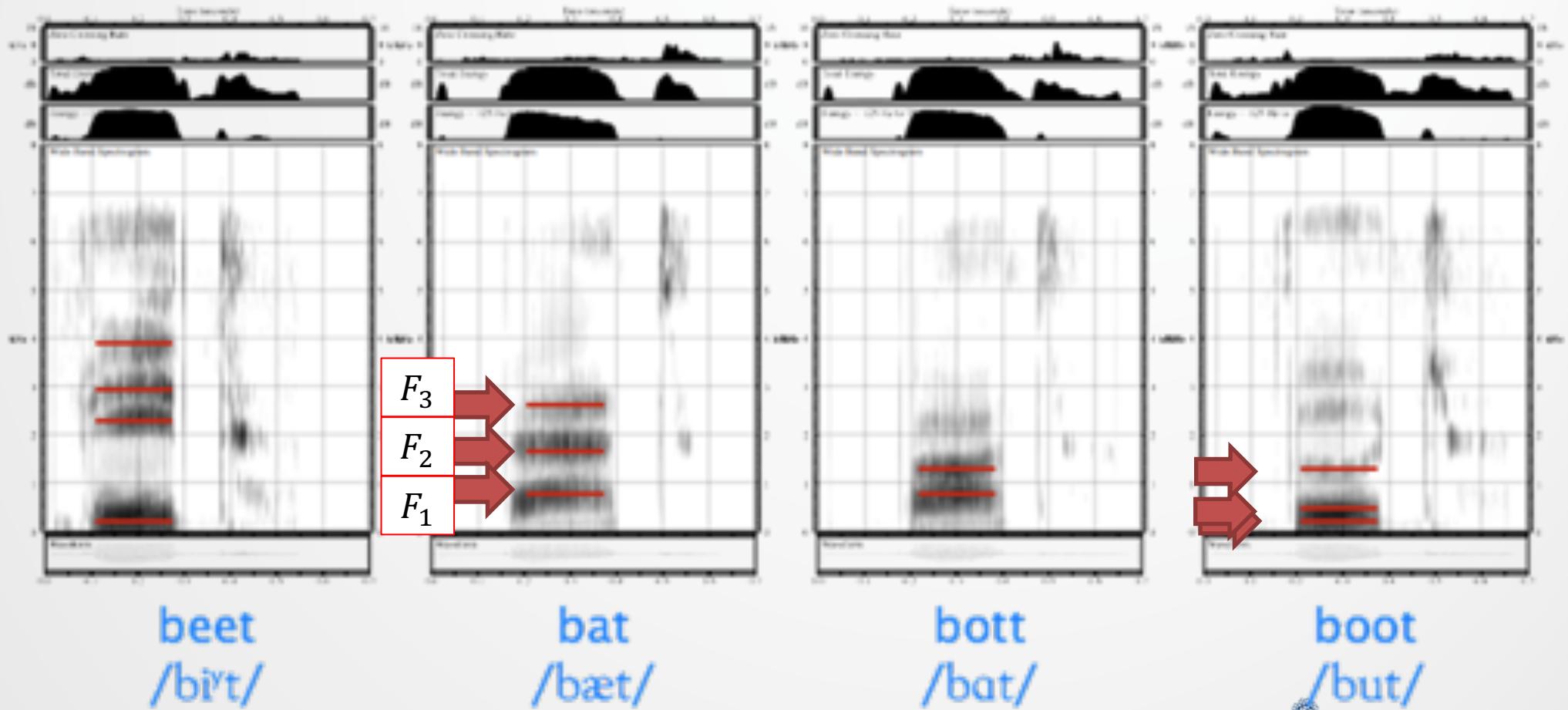
Waves in concatenated tubes

- Because of partial wave **reflections** that occur at tube boundaries, we can generate spectra with particular **resonances**.

Resonances

UNIVERSITY OF
TORONTO
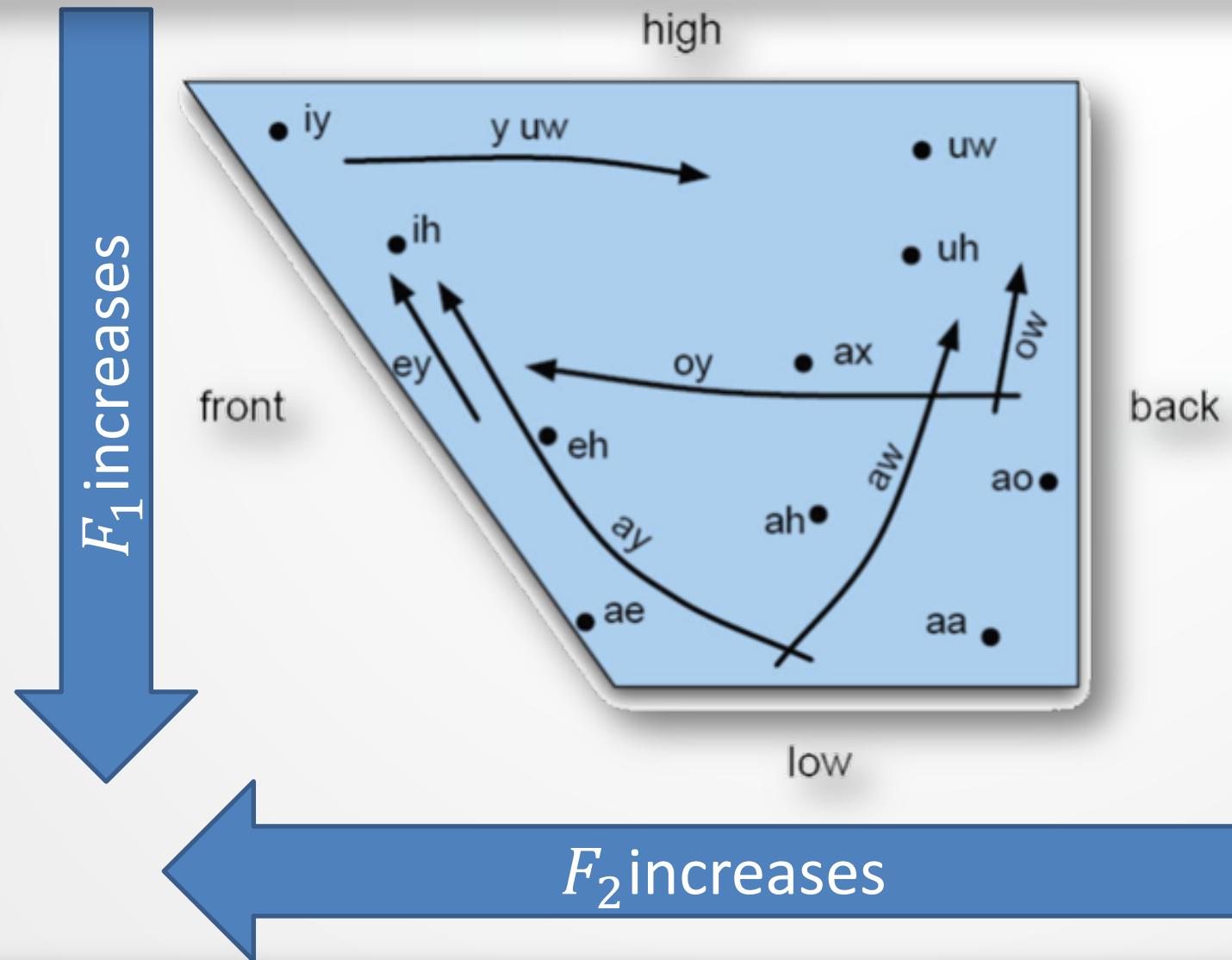
# Formants and vowels
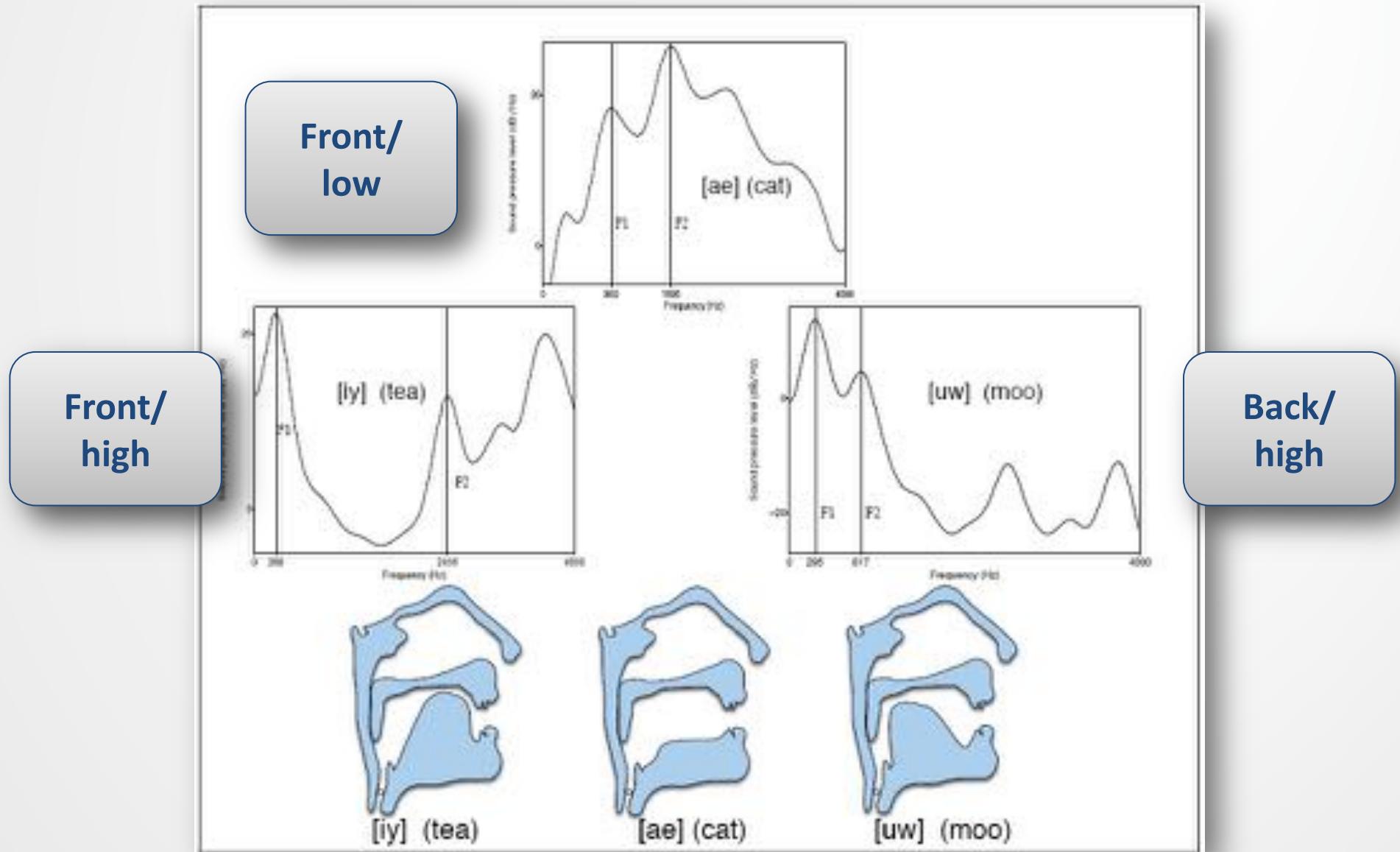
- **Formant**: *n.* A concentration of energy within a frequency band. Ordered from low to high bands (e.g., F_1, F_2, F_3).

beet /biˠt/

bat /bæt/

bott /bɑt/

boot /but/

UNIVERSITY OF TORONTO

The vowel trapezoid

Tongues and formants

Front/
low

Front/
high

Back/
high

[iy] (tea)

[ae] (cat)

[uw] (moo)

UNIVERSITY OF
TORONTO
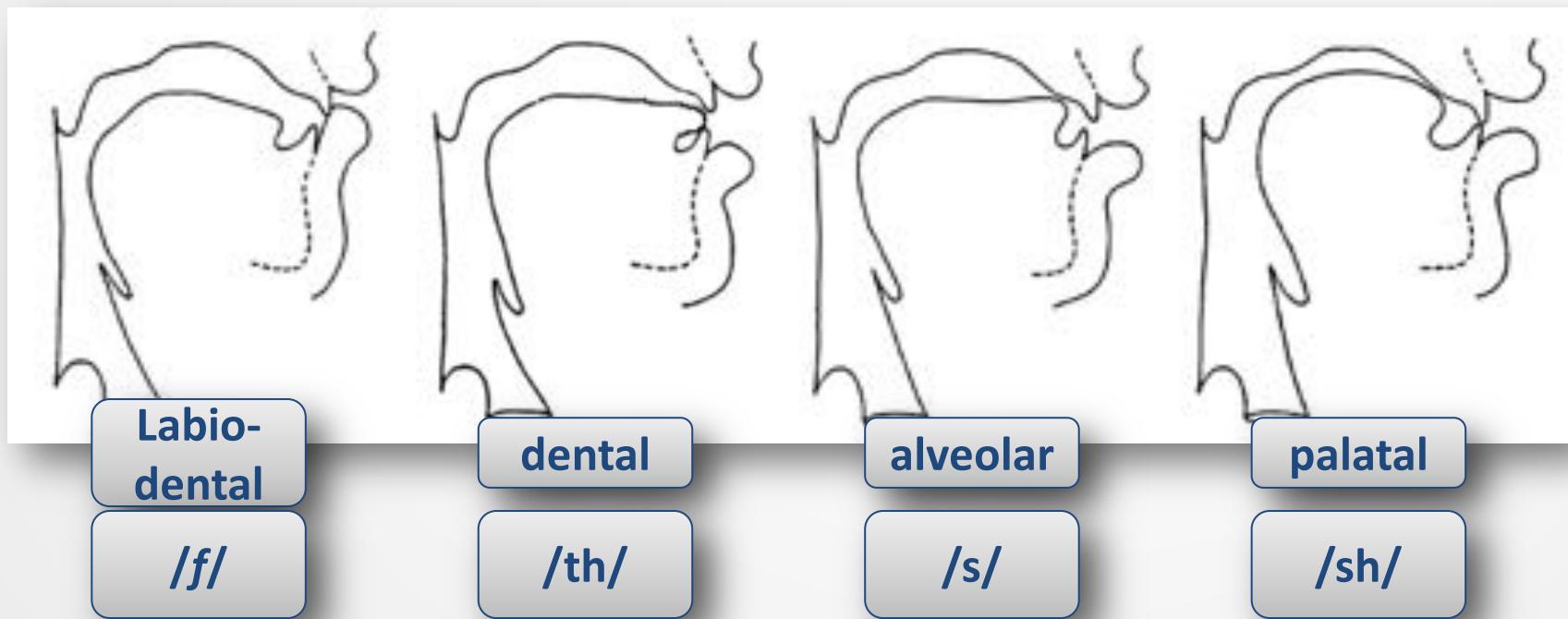
Fricatives (2/6)

- **Fricatives** are caused by acoustic turbulence at a **narrow constriction** whose position determines the sound.

Labio-dental	dental	alveolar	palatal
/f/	/th/	/s/	/sh/

UNIVERSITY OF TORONTO
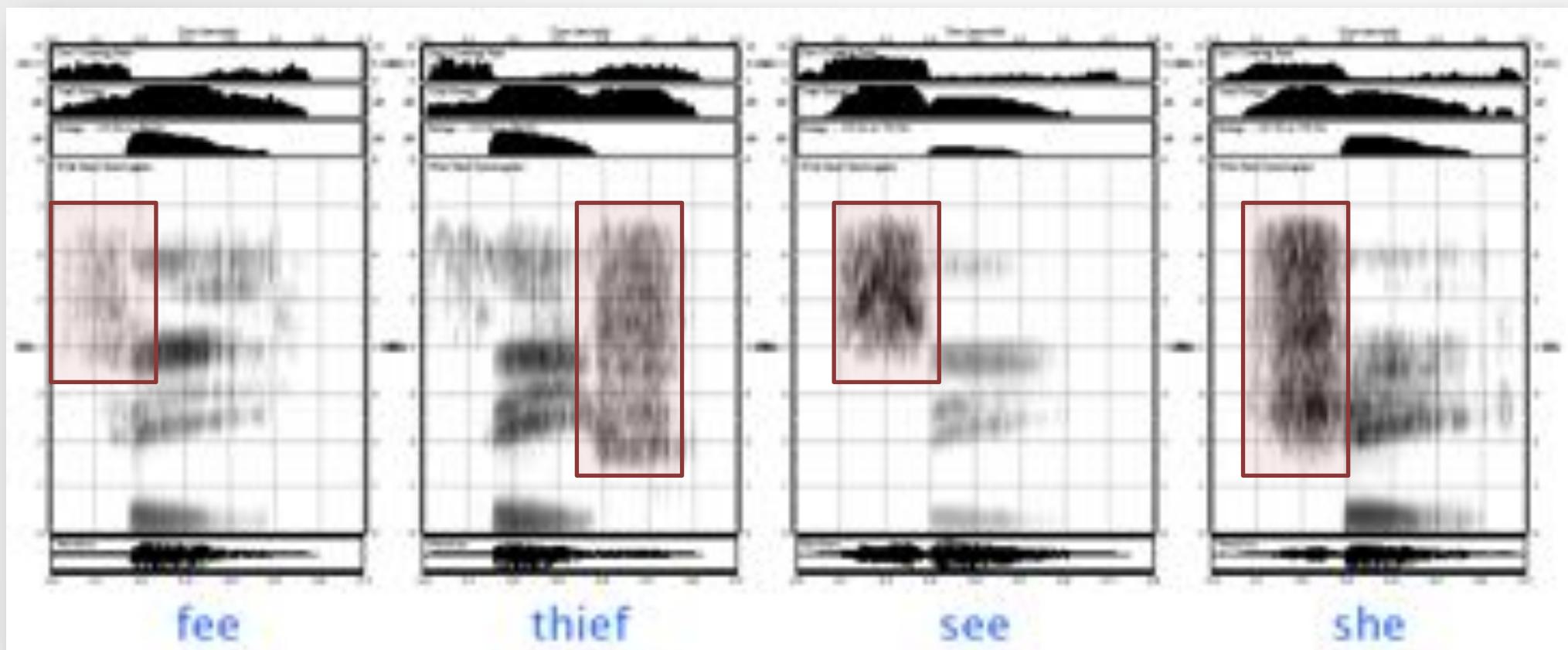
Fricatives

- **Fricatives** have four places of articulation.
- Each place of articulation has a **voiced** fricative (i.e., the glottis can be vibrating), and an **unvoiced** fricative.
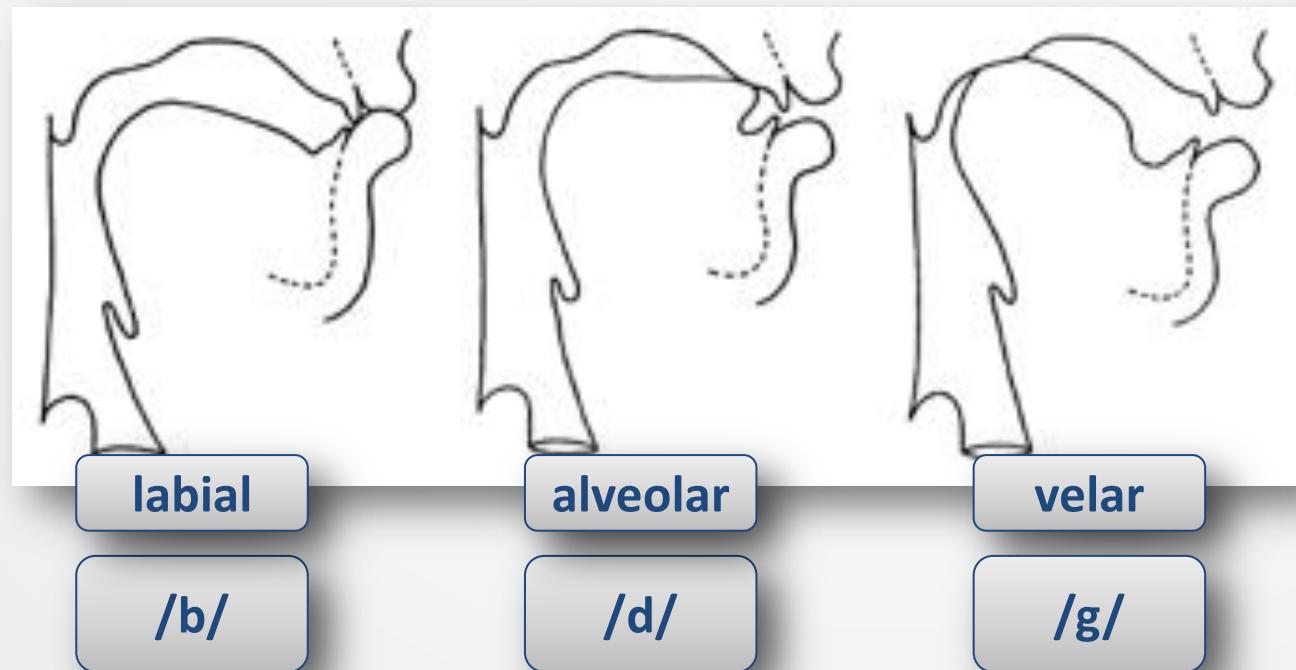
	Unvoiced		Voiced	
Labio-dental	/f/	_**f**ee_	/v/	_**V**endetta_
Dental	/th/	_**th**ief_	/dh/	_**Th**ee_
Alveolar	/s/	_**s**ee_	/z/	_**Z**ardo**z**_
Palatal	/sh/	_**sh**e_	/zh/	_**Zh**a-**zh**a_

UNIVERSITY OF TORONTO

Unvoiced fricatives

fee   thief   see   she

UNIVERSITY OF TORONTO

Plosives (3/6)

- **Plosives** build pressure behind a **complete closure** in the vocal tract.
- A **sudden release** of this constriction results in **brief noise**.

labial	alveolar	velar
/b/	/d/	/g/

UNIVERSITY OF TORONTO
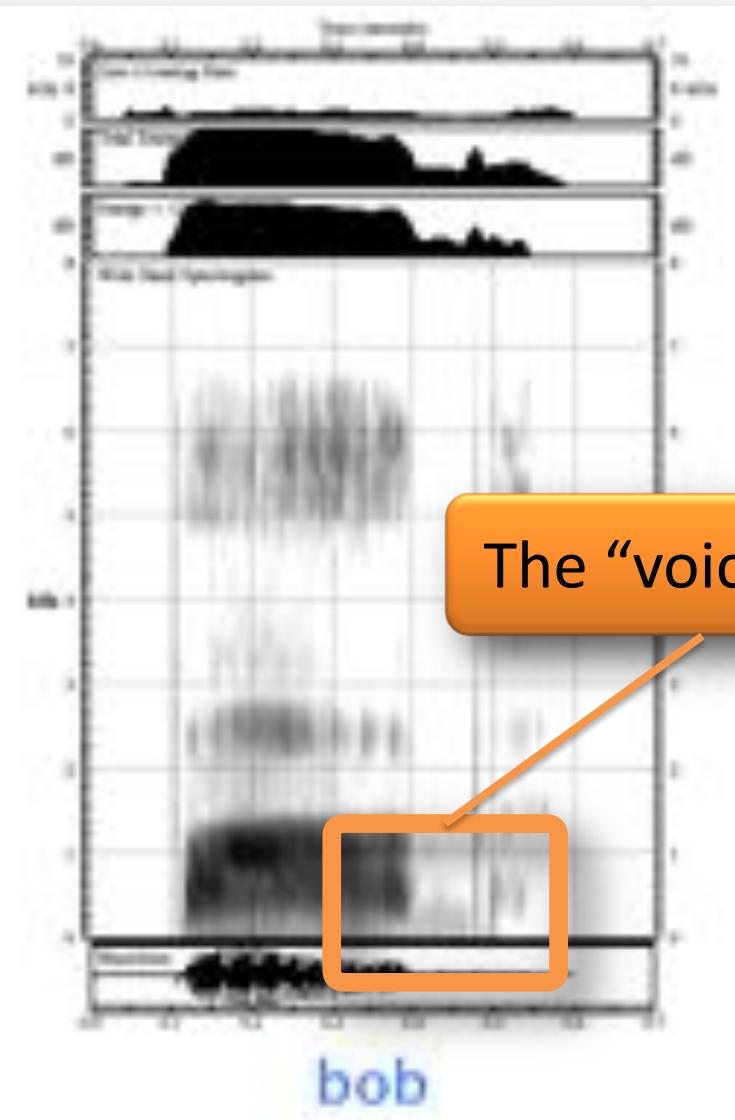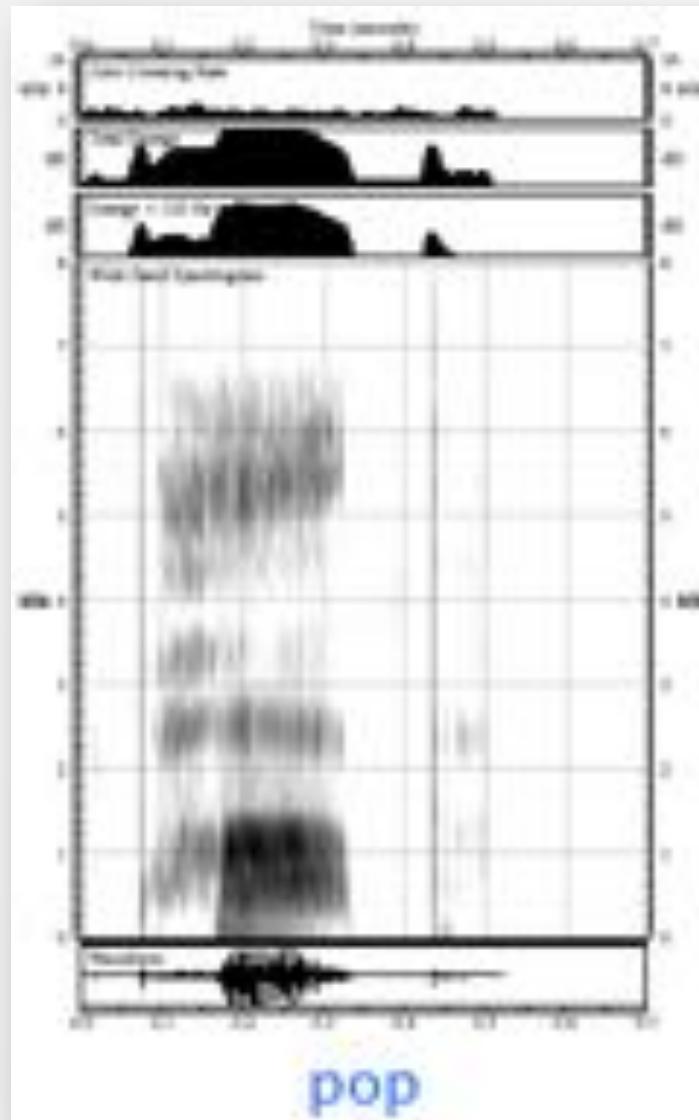
Plosives

- **Plosives** have three places of articulation:
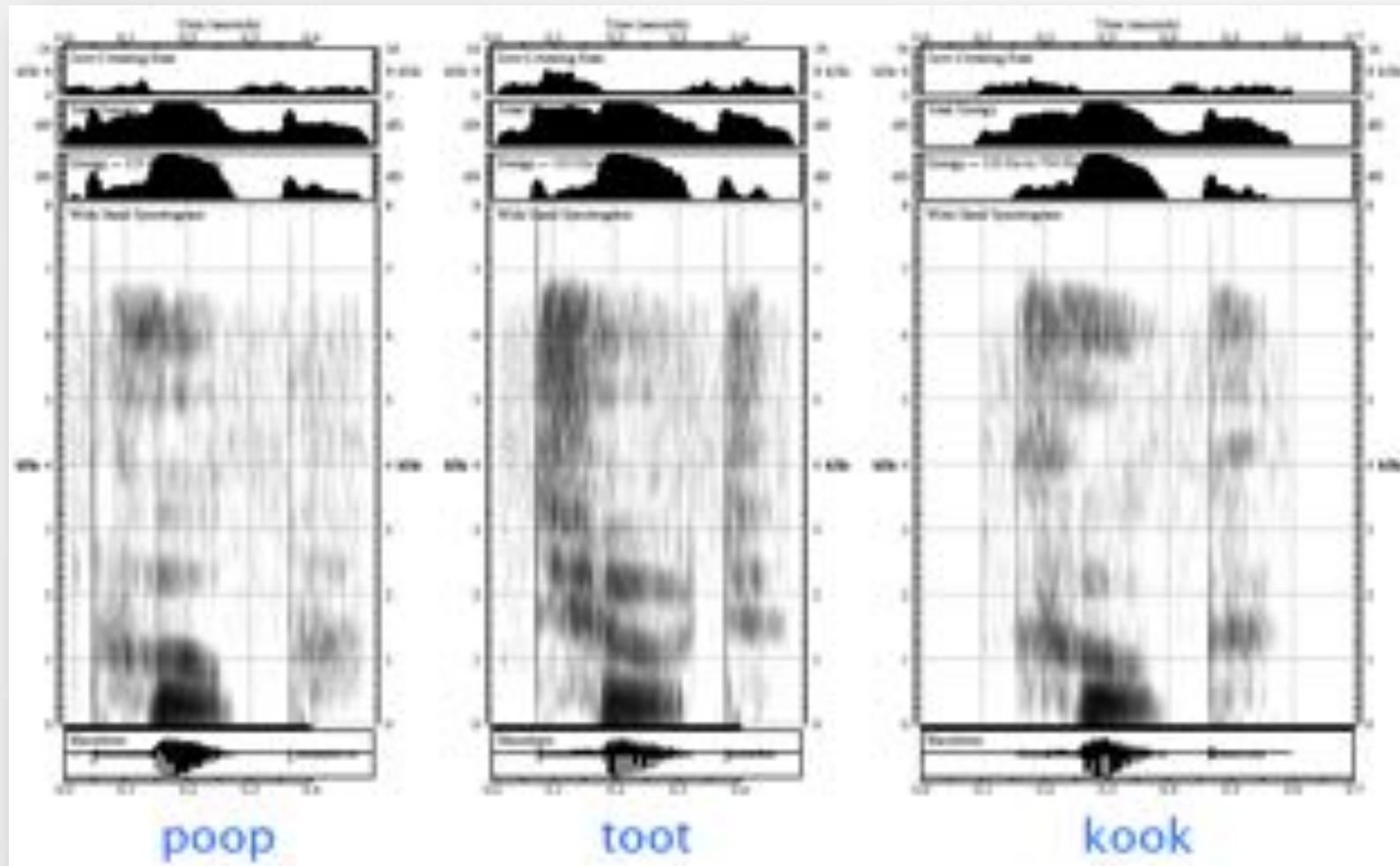
	Unvoiced		Voiced	
Labial	/p/	_**p**or**p**oise_	/b/	_**b**a**b**oon_
Alveolar	/t/	_**t**or**t**_	/d/	_**d**o**d**o_
Velar	/k/	_**k**i**ck**_	/g/	_**G**oo**g**le_

- **Voiced** stops are usually characterized by a "**voice bar**" during closure, indicating the vibrating glottis.
- Formant **transitions** are very **informative** in classification.

UNIVERSITY OF TORONTO

Voicing in plosives

The "voice bar"

pop

bob

UNIVERSITY OF
TORONTO
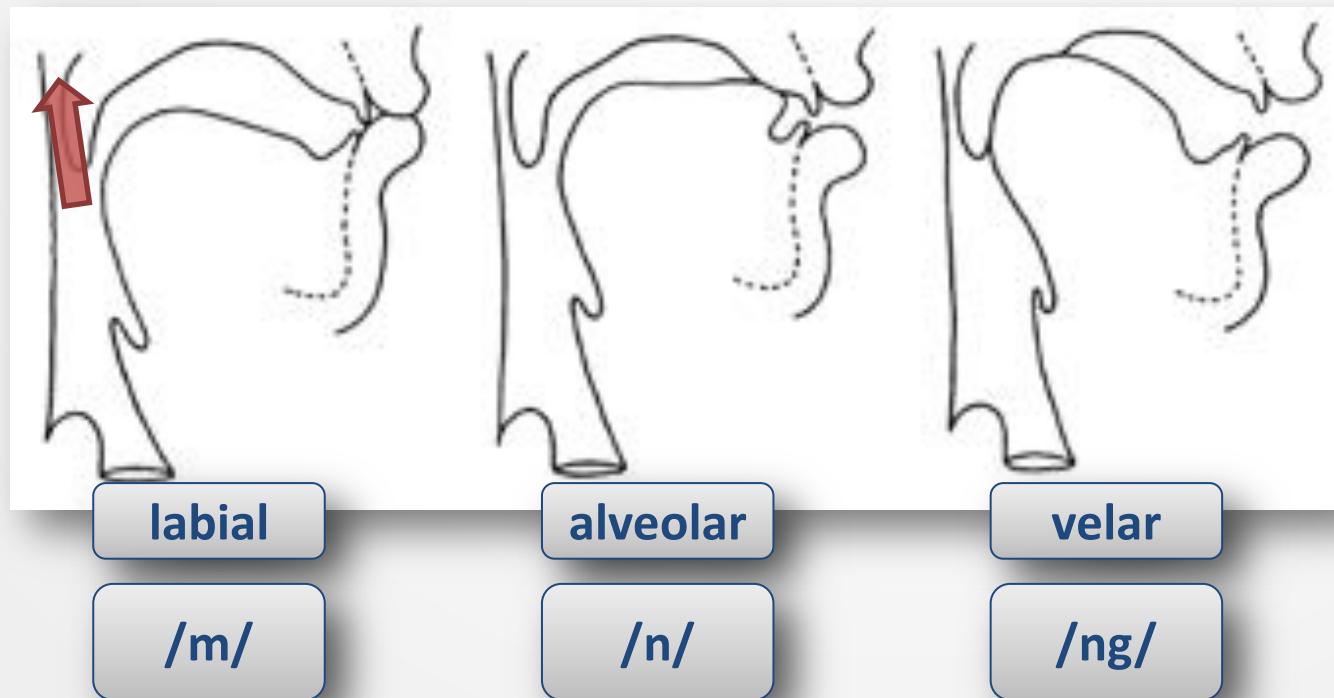
Formant transitions in plosives

poop　　　　　　toot　　　　　　kook
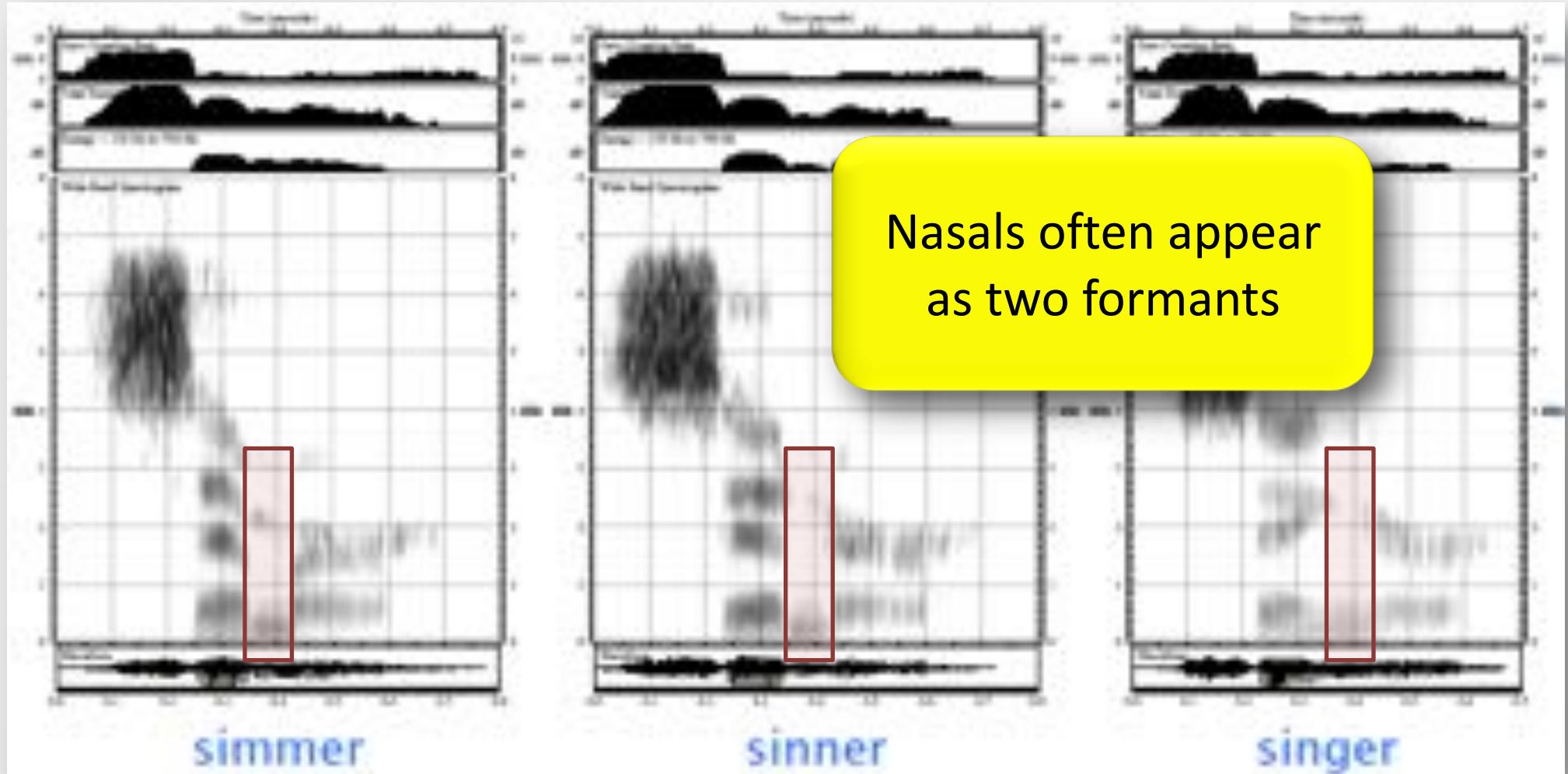
- Despite a **common** vowel, the **motion** of F_2 and F_3 into (and out of) the vowel helps identify the plosive.

UNIVERSITY OF TORONTO

Nasals (4/6)

- **Nasals** involve lowering the velum so that air passes through the **nasal cavity**.
- **Closures** in the oral cavity (at same positions as plosives) change the resonant characteristics of the nasal sonorant.

labial	alveolar	velar
/m/	/n/	/ng/
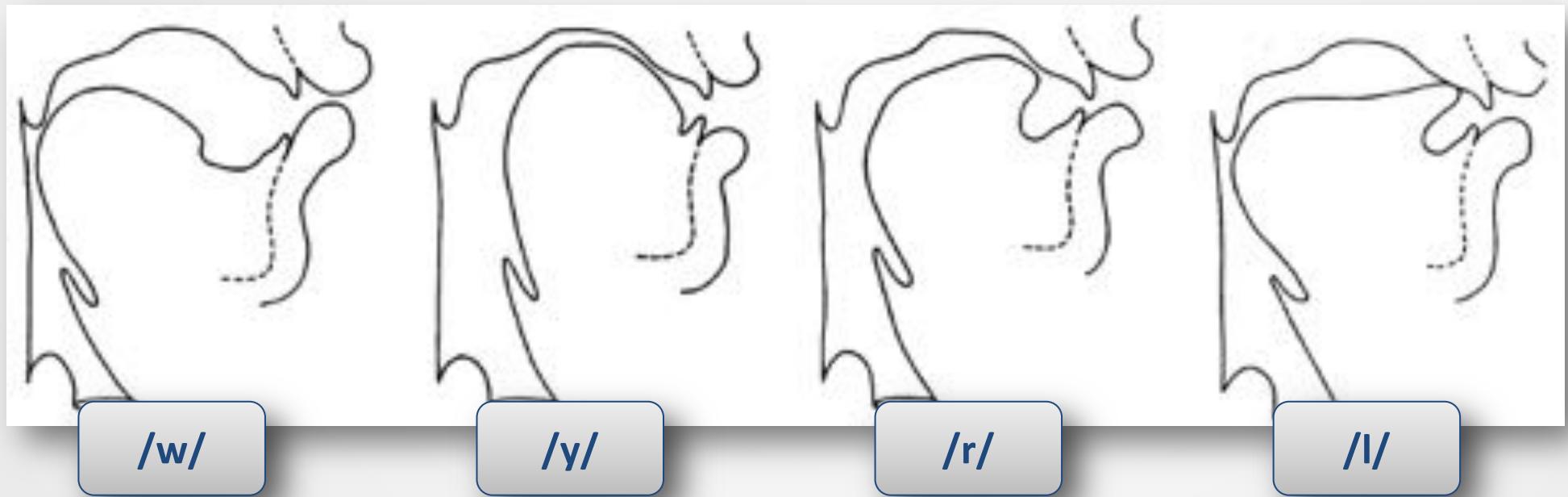
Formant transitions among nasals

Nasals often appear as two formants

simmer   sinner   singer

- Despite a common vowel, the motion of F_2 and F_3 before and after each nasal helps to identify it.

UNIVERSITY OF TORONTO

Semivowels (5/6)

- **Semivowels** act as consonants in syllables and involve constriction in the vocal tract, but there is **less turbulence**.
  - They also involve slower articulatory motion.
- **Laterals** involve airflow around the **sides** of the tongue.

| /w/ | /y/ | /r/ | /l/ |

UNIVERSITY OF TORONTO
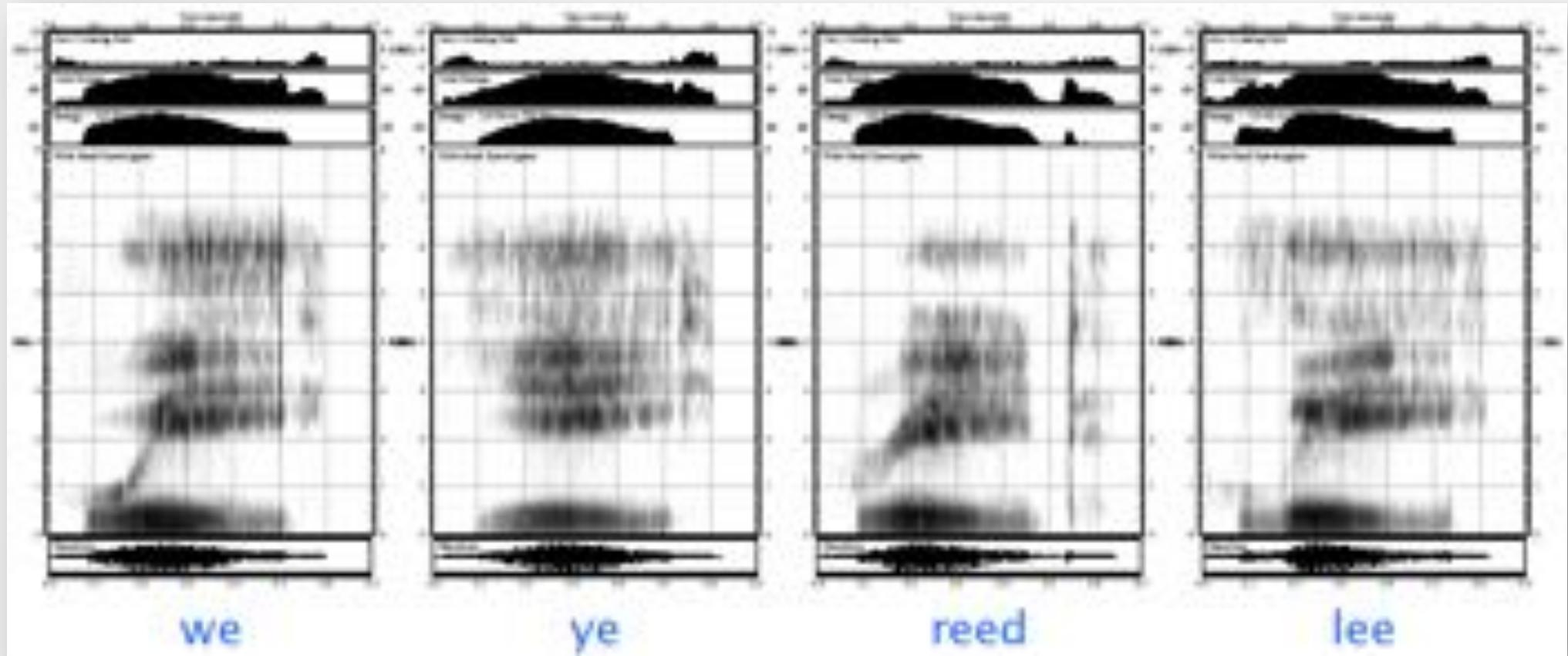
Semivowels

- Semivowels are often sub-classified as glides or liquids.

	Semivowel		Nearest vowel
Glides	/w/	_**W**o**w**_	/uw/
	/y/	_**y**o**y**o_	/iy/
Liquids	/r/	_**r**ea**r**_	/er/
	/l/	_**L**u**l**u_	/ow/

- Semivowels are more constricted versions of corresponding vowels.
  - Similar formants, though generally weaker.

UNIVERSITY OF TORONTO

Semivowels

we   ye   reed   lee
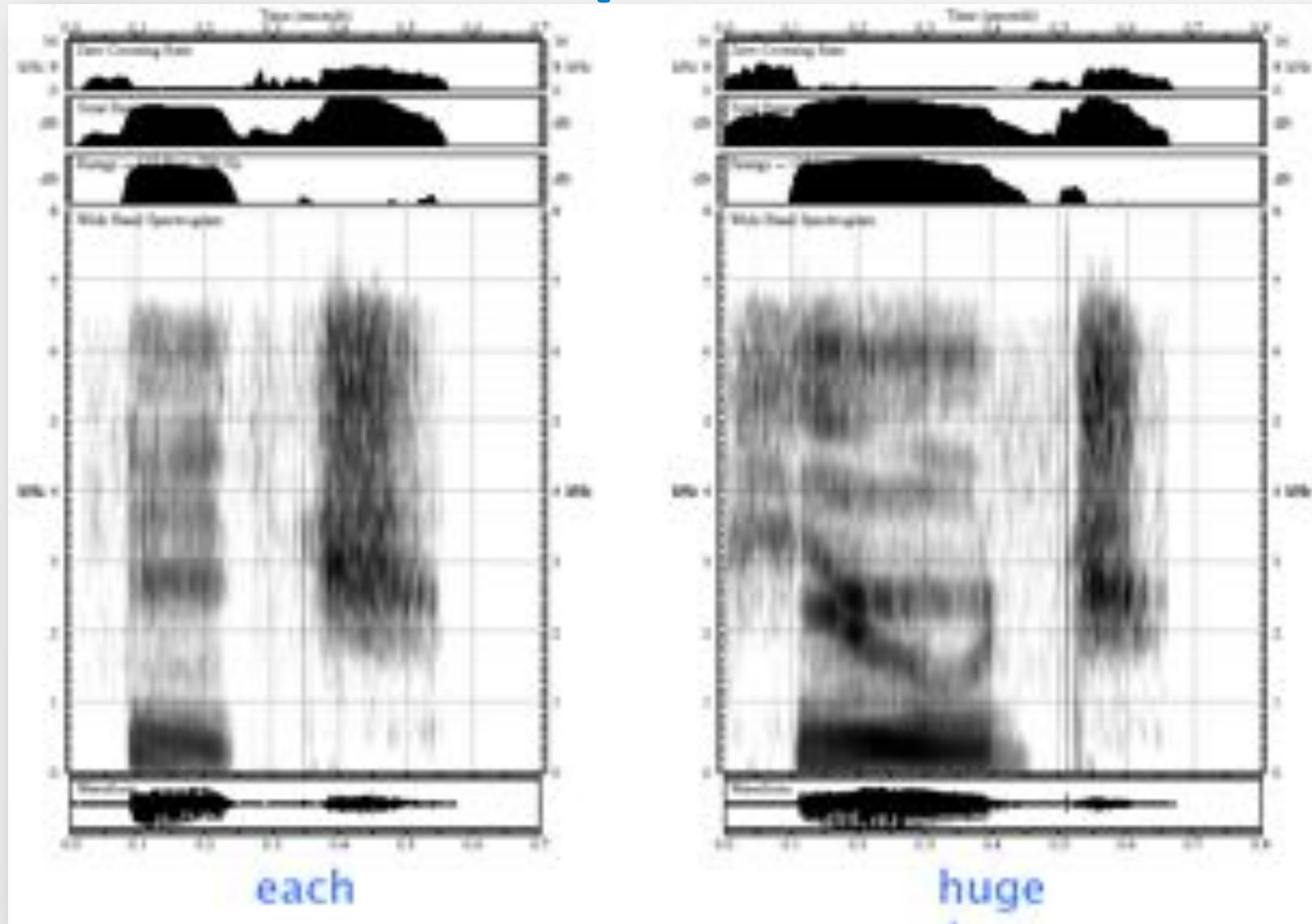
- Note the drastic formant transitions which are more typical of semivowels.

UNIVERSITY OF
TORONTO

Affricates and aspirants (6/6)

- There are two **affricates**: /jh/ (voiced; e.g., *judge*) and /ch/ (unvoiced; e.g., *church*).
  - These involve an **alveolar stop** followed by a **fricative**.
  - Voicing in /jh/ is normally indicated by voice bars, as with plosives.

- There's only one **aspirant** in Canadian English: /h/ (e.g., *hat*)
  - This involves turbulence generated at the **glottis**,
  - In Canadian English, there is **no** constriction in the vocal tract.

Affricates and aspirants

each

huge

Alternative pronunciations

- **Pronunciations** of words can vary significantly, but with observable **frequencies**.
  - The **Switchboard** corpus is a phonetically annotated database of speech recorded in telephone conversations.
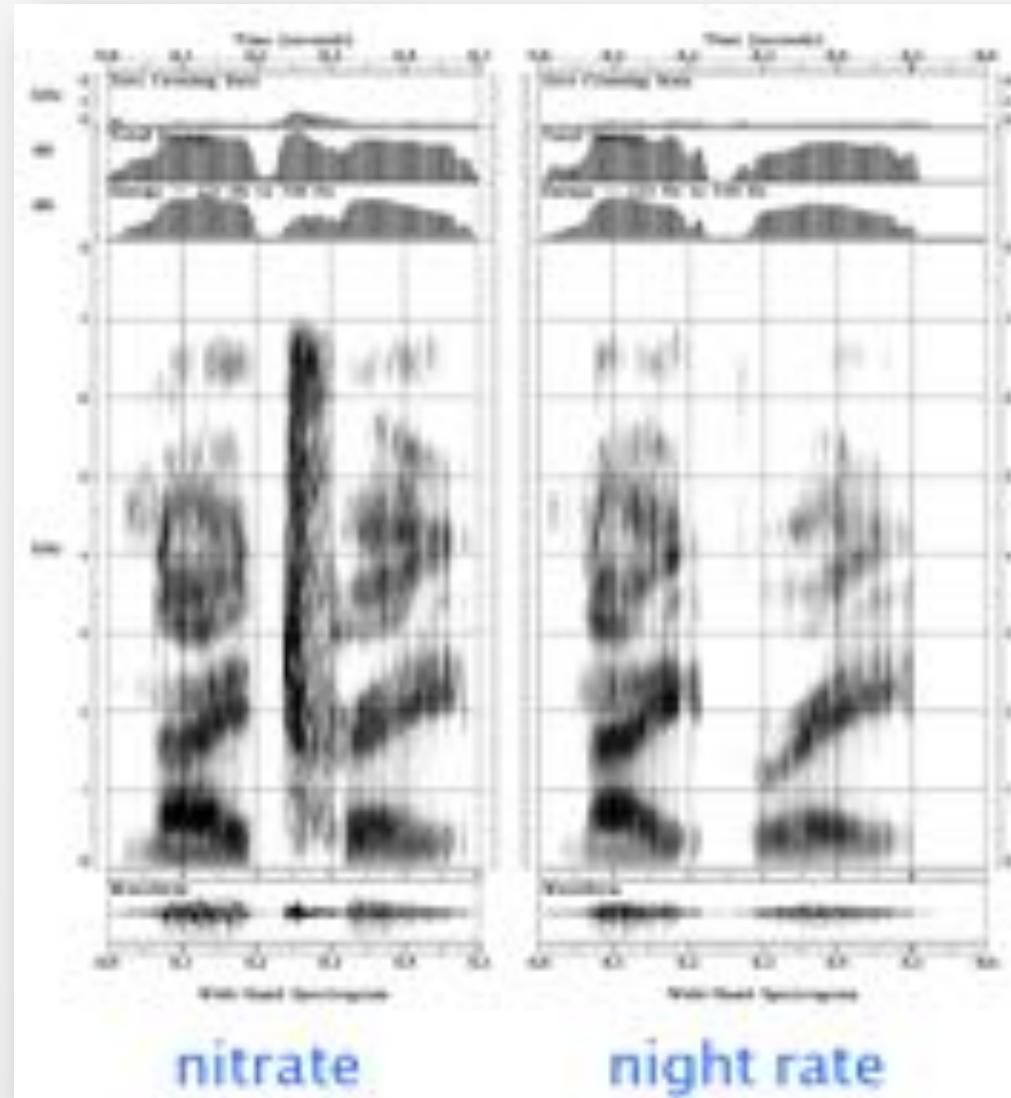
because				about			
ARPAbet	%	ARPAbet	%	ARPAbet	%	ARPAbet	%
b iy k ah z	27%	k s	2%	ax b aw	32%	b ae	3%
b ix k ah z	14%	k ix z	2%	ax b aw t	16%	b aw t	3%
k ah z	7%	k ih z	2%	b aw	9%	ax b aw dx	3%
k ax z	5%	b iy k ah zh	2%	ix b aw	8%	ax b ae	3%
b ix k ax z	4%	b iy k ah s	2%	ix b aw t	5%	b aa	3%
b ih k ah z	3%	b iy k ah	2%	ix b ae	4%	b ae dx	3%
b ax k ah z	3%	b iy k aa z	2%	ax b ae dx	3%	ix b aw dx	2%
k uh z	2%	ax z	2%	b aw dx	3%	ix b aa t	2%

Known effects of pronunciation

- Speakers tend to **drop** or **change** pronunciations in **predictable** ways in order to reduce the effort required to **co-ordinate** the various articulators.
  - **Palatalization** generally refers to a **conflation** of phonemes closer to the frontal palate than they 'should' be.
  - **Final t/d deletion** is simply the **omission** of alveolar plosives from the ends of words.
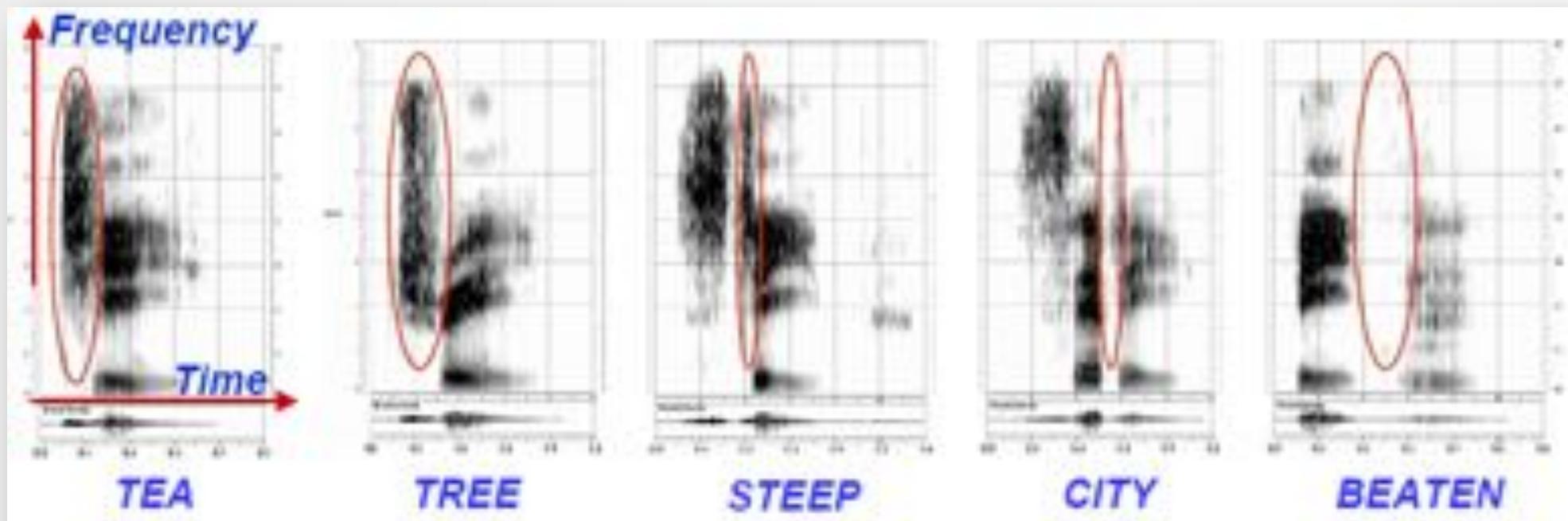
| Palatalization			Final t/d Deletion		
Phrase	**Lexical**	**Reduced**	**Phrase**	**Lexical**	**Reduced**
set your	s eh t y ow r	s eh ch er	find him	f ay n d h ih m	f ay n ix m
not yet	n aa t y eh t	n aa ch eh t	and we	ae n d w iy	eh n w iy
did you	d ih d y uw	d ih jh y ah	draft the	d r ae f t dh iy	d r ae f dh iy

Variation at syllable boundaries

nitrate   night rate

UNIVERSITY OF
TORONTO
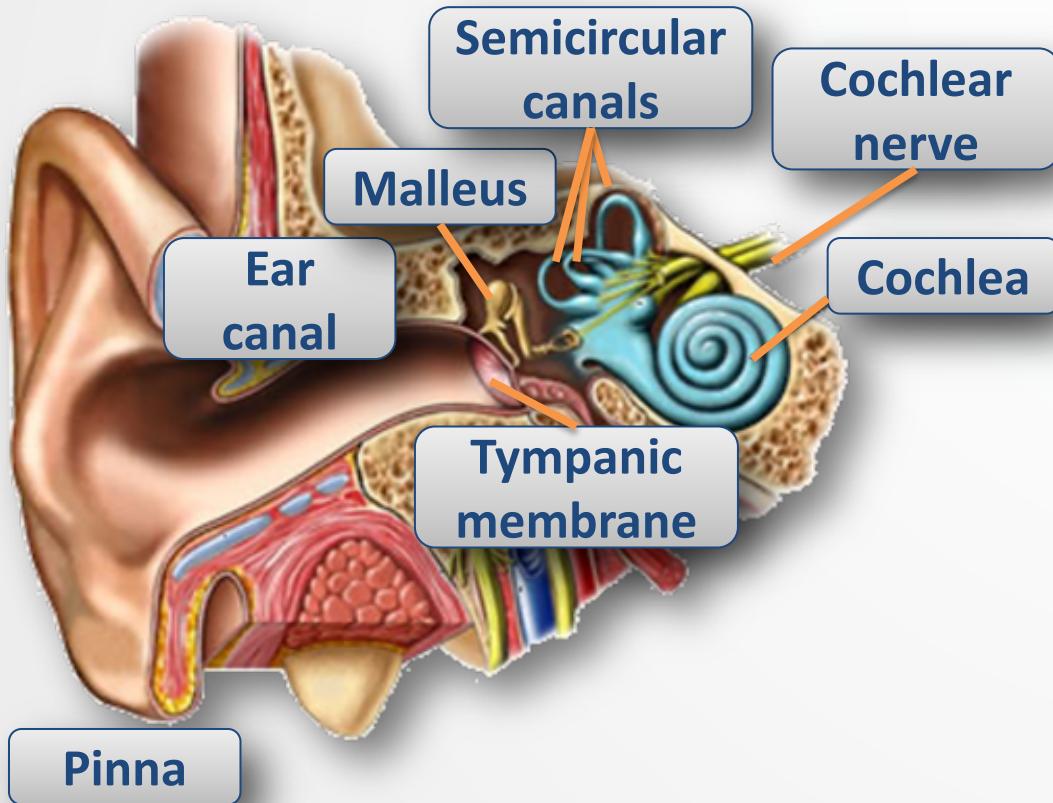
Phonological variation

- The acoustics of a phoneme depend strongly on the **context** in which that phoneme occurs.
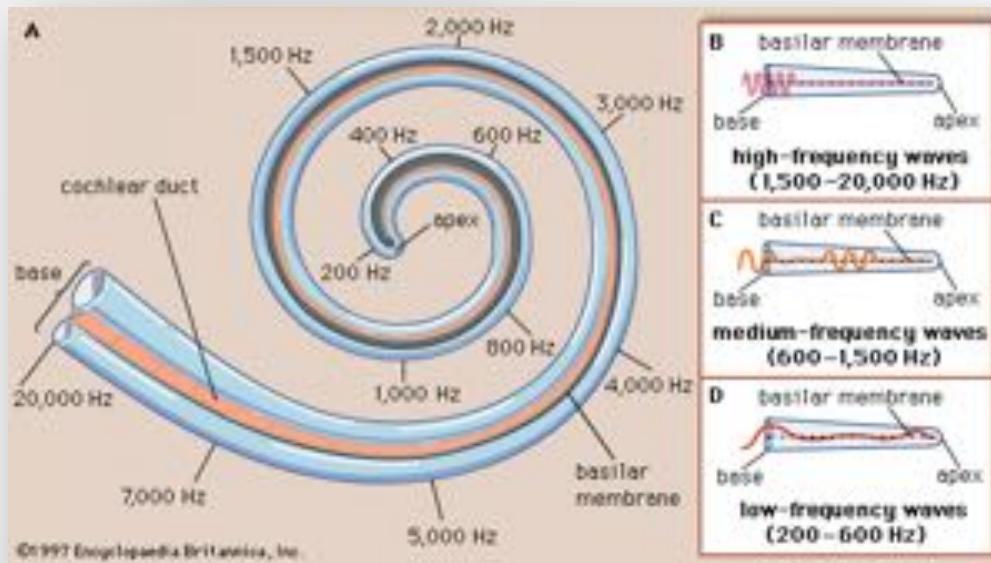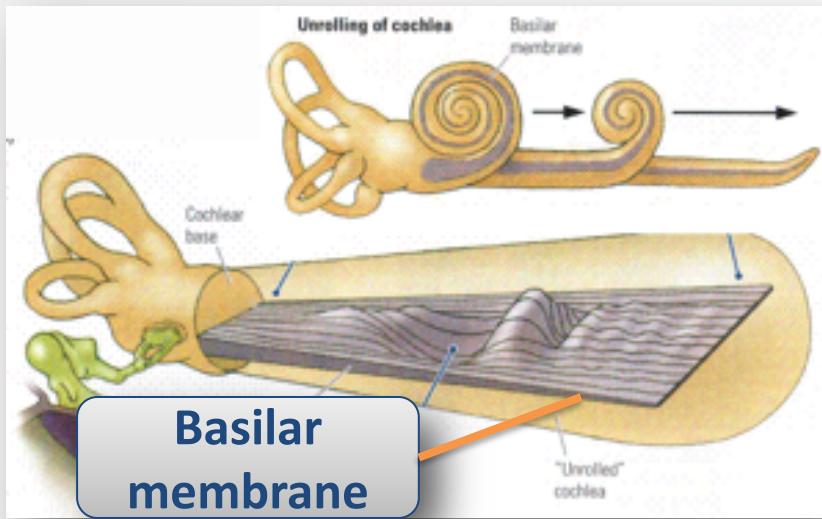
*That must make **recognizing** phonemes hard, right?*
*How do humans do it?*

UNIVERSITY OF TORONTO

The inner ear

- Time-variant waves enter the ear, vibrating the **tympanic membrane**.

- This membrane causes tiny bones (incl. **malleus**) to vibrate.

- These bones in turn vibrate a structure within a shell-shaped bony structure called the **cochlea**.

UNIVERSITY OF TORONTO

The cochlea and basilar membrane
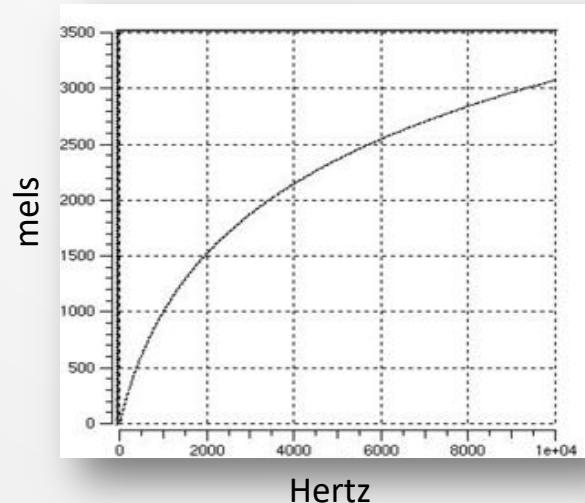
Basilar membrane

- The **basilar membrane** is covered with tiny hair-like nerves – some near the **base**, some near the **apex**.

- **High** frequencies are picked up near the base, **low** frequencies near the apex.

- These nerves fire when activated, and communicate to the brain.

UNIVERSITY OF TORONTO

The Mel-scale

- Human hearing is **not** equally sensitive to **all** frequencies.
  - We are **less** sensitive to frequencies > 1 kHz.

- A **mel** is a unit of pitch. Pairs of sounds which are **perceptually** equidistant in pitch are separated by an equal number of **mels**.
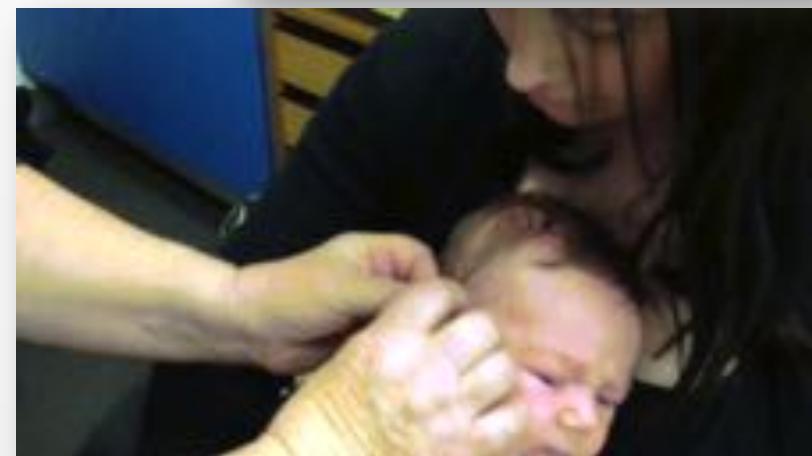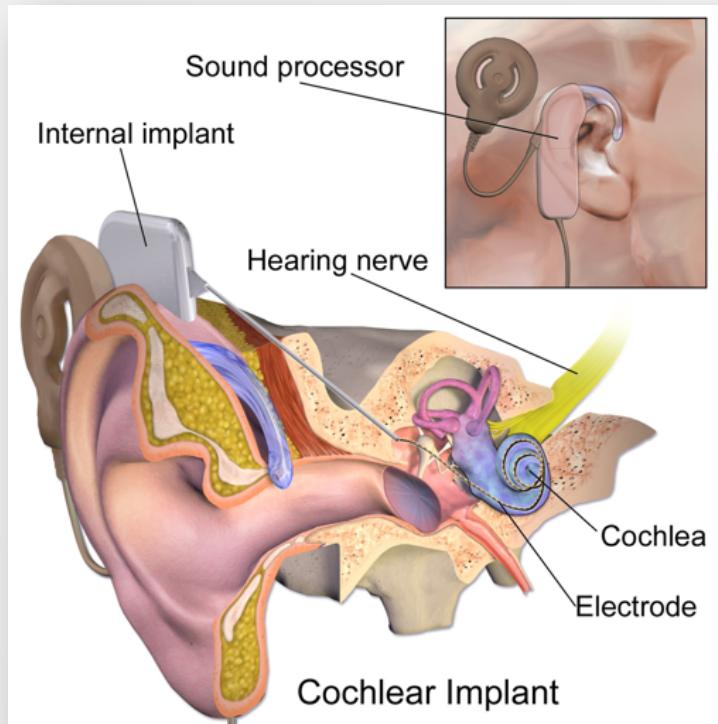
$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

(No need to memorize this either)

UNIVERSITY OF TORONTO

Aside – Challenges of perception

- **Cochlear implants** replace the basilar membrane and stimulate the auditory nerve directly.

Sound processor
Internal implant
Hearing nerve
Cochlea
Electrode
Cochlear Implant

UNIVERSITY OF TORONTO

Next…

- How the Mel scale is used in ASR.

- Automatic speech recognition.

UNIVERSITY OF TORONTO